The genetics of oesophageal cancer in South African populations

Bye, Hannah

Awarding institution:
King's College London

The genetics of oesophageal cancer in South African populations

Hannah Bye

A thesis submitted to King's College London in fulfilment of the degree of Doctor of Philosophy

Department of Medical and Molecular Genetics
Division of Genetics and Molecular Medicine
King's College London

November 2013

Abstract

Oesophageal cancer is a common form of cancer in South Africa. The predominant subtype is oesophageal squamous cell carcinoma (OSCC) and the disease has a very poor prognosis. The aim of this thesis was to investigate the genetics of OSCC in the South African Black and Mixed Ancestry populations.

Genetic susceptibility to OSCC was explored initially through case-control association studies of 18 variants with strong evidence of association with the disease in other populations. Most loci did not show association in the South African populations. However, in the Mixed Ancestry population, *ALDH2* +82 A>G (rs886205) and *RUNX1* rs2014300 were associated with OSCC (odds ratio (OR) = 0.70, 95% confidence interval (CI) = 0.55-0.89, P=0.0038; and OR = 1.33, 95% CI = 1.09-1.63, P=0.0055, respectively). Further investigation of *PLCE1* in the Black population found Arg548Leu (rs17417407) to be associated with disease (OR = 0.74, 95% CI = 0.60-0.93; P=0.008). These findings suggest that there may be substantial differences in the genetic architecture of OSCC in African populations. Additionally, genetic susceptibility was explored in the Black population using the Immunochip, a genotyping array containing ~200,000 variants. Although no polymorphisms were significantly associated with OSCC, several variants in *TGFBR3* showed suggestive evidence of association, which was promising as the TGF-β pathway has been shown to have an important role in the development of the disease.

In a pilot study to investigate somatic mutations in OSCC, the whole-exomes of 8 blood-tumour pairs were sequenced, with mutations identified in several tumour suppressor genes, including *TP53, KL* and *APC*. Sanger sequencing of two candidate genes, *TP53* and *PPM1D*, in all available samples, showed that 60% of tumours contained *TP53* mutations, and that 36% of tumours showed evidence of loss of heterozygosity at the *PPM1D* locus, suggesting that it may be an important alteration for OSCC development.

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Chris Mathew, for giving me the opportunity to undertake this project, and for providing me with guidance and his expertise throughout my time here. Thank you also to my second supervisor, Professor Cathryn Lewis, for her valued input into the statistical analysis aspects of the project.

Many thanks to Dr Michael Simpson and Dr Sarah Spain, whose bioinformatics skills, in NGS sequencing and Immunochip data, respectively, were invaluable and provided me with data in a format I could work with.

I also thank and acknowledge the staff at the BRC Genomics Core facility for processing samples on the Immunochip arrays and the next generation sequencers.

I am grateful to Dr Iwanka Kozarewa, Dr James Campbell and their other colleagues at the Institute of Cancer Research, London, with whom we collaborated with on the exome sequencing project. They were generous in sharing their protocols with us and allowed me to spend several days in their lab observing the work, as well as carrying out the sequencing and analysis for some of our samples.

Thanks also to Professor Iqbal Parker and his oesophageal cancer research group at the University of Cape Town, South Africa, who provide us with samples from their local populations. I am grateful to the volunteers, both cancer patients and controls, who allow a blood sample or tumour tissue to be used for research purposes.

I would also like to thank the rest of the Complex Disease Group, past and present, Dr Natalie Prescott, Dr Kirstin Taylor, Kristina Stone and Dr Kathy

Dominy, for the scientific discussions and advice given throughout my project, as well as providing an enjoyable working environment.

I am grateful to the Medical Research Council (MRC) and The Generation Trust for funding my PhD. In addition, the oesophageal cancer research project has benefitted from funding from the Association for International Cancer Research (AICR).

Lastly, I would like to thank my family. My mum, Christine, has been dedicated to my upbringing and development throughout my life, and thanks to her, I have been able to succeed at school and university to be where I am today. Thank you, mum, I appreciate everything you have done for me and I dedicate this thesis to you. Thanks also to my brother, Daniel, who set high educational standards for me to follow throughout childhood. Finally, thank you to my boyfriend, Dan, who listens to the daily updates of my work and has given his advice on the statistical aspects of my thesis, for which I am grateful.

This thesis is dedicated to my mum, Christine.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ASW	African ancestry in Southwest USA
bp	Base pair
BMI	Body mass index
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CGH	Comparative genome hybridization
CI	Confidence interval
COSMIC	Catalogue of Somatic Mutations in Cancer
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
GERD	Gastroesophageal reflux disease
GWAS	Genome-wide association study
HIV	Human immunodeficiency virus
HNSCC	Head and neck squamous cell carcinoma
HPV	Human papillomavirus
HWE	Hardy Weinberg Equilibrium
IARC	International Agency for Research on Cancer
ICGC	International Cancer Genome Consortium
ICR	Institute of Cancer Research
IGV	Integrative Genomics Viewer
JPT	Japanese in Tokyo, Japan
KCL	King's College London
LD	Linkage disequilibrium
LOH	Loss-of-heterozygosity
MAF	Minor allele frequency
Mb	Mega base
mRNA	Messenger ribonucleic acid
NSAIDS	Non-steroidal anti-inflammatory drugs
OAC	Oesophageal adenocarcinoma
OR	Odds ratio

ORF	Open reading frame
OSCC	Oesophageal squamous cell carcinoma
PCA	Principal Components Analysis
PCR	Polymerase chain reaction
QC	Quality control
rpm	Revolutions per minute
SNP	Single nucleotide polymorphism
TCGA	The Cancer Genome Atlas
UTR	Untranslated region
UV	Ultraviolet
WTCCC	Wellcome Trust Case Control Consortium
YRI	Yoruba in Ibadan, Nigeria

1 Introduction

Cancer is the third leading cause of death world-wide after cardiovascular disease and infectious and parasitic diseases (World Health Organization 2008), with approximately 12.7 million cases diagnosed and 7.6 million deaths occurring world-wide in 2008 (Ferlay *et al.* 2010.a). In the USA, this was estimated to cost $124.6 billion in 2010 (National Cancer Institute 2012). This is likely to increase in the future, with more cancers occurring due to an expanding and aging population. The need to reduce costs and save lives through prevention or early detection is, therefore, essential. Two approaches are vital to this: avoidance of known environmental risk factors and the application of personalized medicine. The latter would enable cancer treatment to be specific to each patient, with their inherited genome and their cancer genome guiding treatment. This approach is still a long way from being common practice in the clinic, with the genetics of cancer needed to be further explored. This includes identifying both genetic variants which affect cancer susceptibility and the somatic changes that occur in the tumour.

Certain cancers are better understood than others, with those common in the industrialized world, such as breast and colorectal cancer, receiving the most funding for research. Oesophageal squamous cell carcinoma is common in the developing world with a high incidence in regions of South Africa and Asia. This thesis aims to improve our knowledge of both the genetic susceptibility and the somatic mutations that occur in oesophageal squamous cell carcinoma in the South African Black and Mixed Ancestry populations.

1.1 Oesophageal cancer

The oesophagus is a muscular tube that connects the pharynx to the stomach, moving food through this region of the digestive tract. It is approximately 18-26 cm in length and has cervical (~6 cm), thoracic (~25 cm) and abdominal (~4 cm) parts (Satvinder and Kang 2008). It is usually referred to in terms of thirds: the upper, middle and distal thirds. The upper region consists of striated muscle, with smooth muscle in the lower section (Kuo and Urma 2006). Oesophageal cancer mainly occurs in the middle and lower thirds.

Oesophageal cancer is the eighth most common form of cancer in the world, with 482,000 newly diagnosed cases in 2008, accounting for 3.8% of all cancers (Ferlay *et al.* 2010.a). This compares to 1.61 million lung cancer cases, which is the cancer with the highest worldwide incidence (12.7% of all cancers). Other cancers with high a incidence include breast (10.9% of all cancers), colorectal (9.7%), stomach (7.8%) and prostate (7.5%). Oesophageal cancer is the sixth most common form of death from cancer, with 407,000 deaths occurring worldwide during 2008 (Ferlay *et al.* 2010.a).

The estimated incidence and mortality rates for oesophageal cancer throughout the world in 2008 are shown in Figure 1.1 (Ferlay *et al.* 2010.a). It is estimated that southern Africa has the highest oesophageal cancer age-standardized incidence (and mortality) rate in the world, with 22.3 cases per 100,000 in males. This is closely followed by males in eastern Asian, with 20.3 cases per 100,000. At the other end of the spectrum, middle and western Africa have the lowest incidence, with approximately 1.5 cases per 100,000 in males. The majority of both cases and deaths occur in developing countries, with incidences in males of 11.8 cases per 100,000 in less developed regions compared to 6.5 cases per 100,000 in more industrialized countries. In all regions, the incidence is greater in males than females.

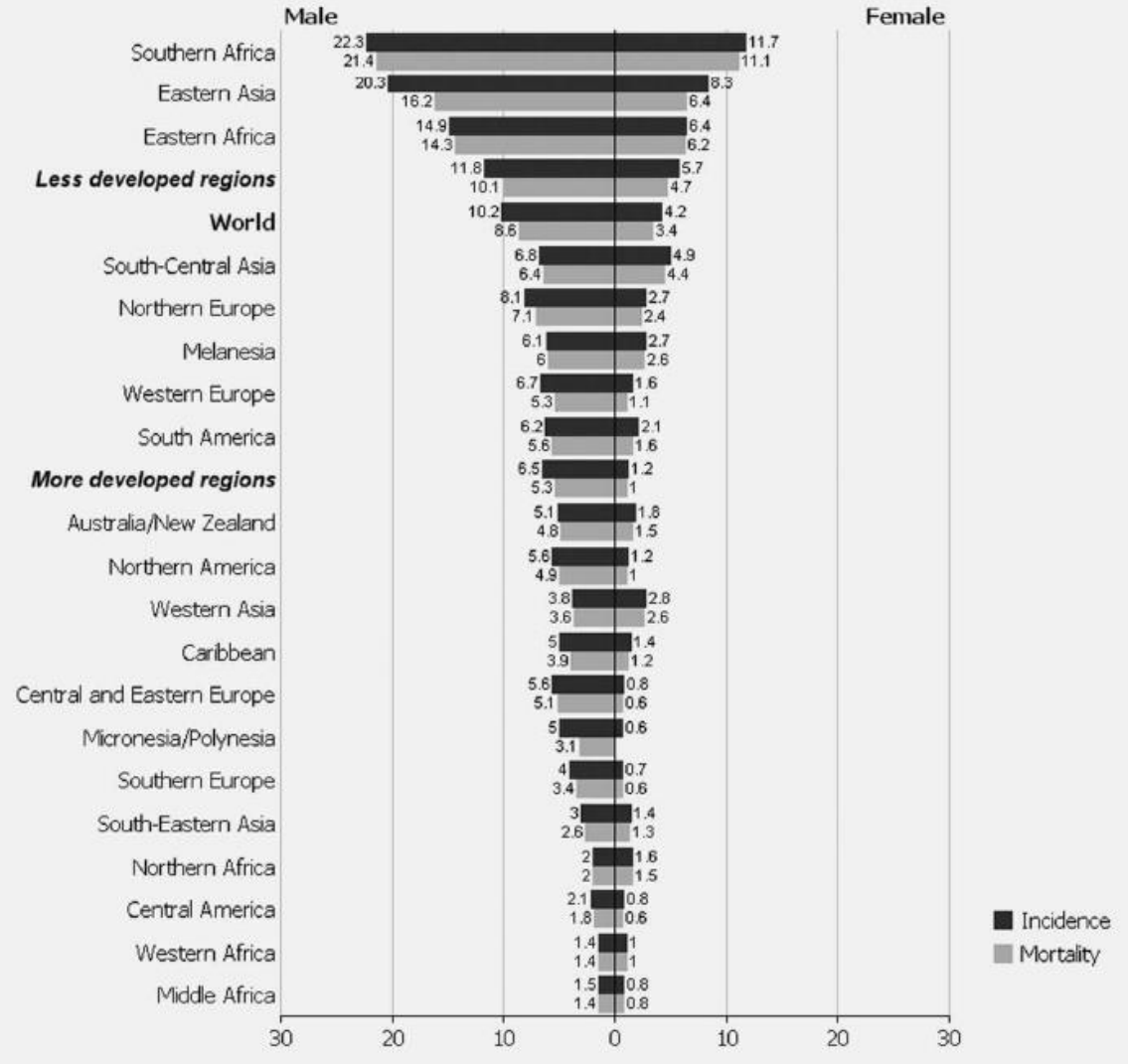


Figure 1.1: Oesophageal cancer incidence and mortality rates
World-wide estimates of age-standardized incidence and mortality rates for oesophageal cancer in 2008 (Ferlay *et al.* 2010.a)

Two main subtypes of oesophageal cancer exist, squamous cell carcinoma and adenocarcinoma, which together account for more than 95% of oesophageal carcinomas (Layke and Lopez 2006). These subtypes are aetiologically unrelated. World-wide, oesophageal squamous cell carcinoma (OSCC) is the most predominant form of oesophageal cancer, mainly occurring in developing countries (Lam 2000). In the western world, OSCC cases have decreased over recent years, and, together with the rise in adenocarcinoma cases, has led to adenocarcinoma being the predominant form in these regions (Holmes and Vaughan 2007).

1.1.1 Oesophageal adenocarcinoma

1.1.1.1 Definition and incidence

Oesophageal adenocarcinoma develops from glandular epithelial cells and predominantly affects the lower third of the oesophagus (Satvinder and Kang 2008). The disease occurs mainly in western countries where the number of cases has increased dramatically over the last four decades and it is now the most common form of oesophageal cancer in these regions (reviewed in Holmes and Vaughan 2007). Worldwide, the UK has the highest incidence of oesophageal adenocarcinoma (Bollschweiler *et al.* 2001). In the USA, the incidence of adenocarcinoma increased from 3.6 cases per million to 25.6 cases per million between 1973 and 2006 (Pohl *et al.* 2010). Incidence rates are showing signs of slowing, with an 1.3% annual increase observed from 1996-2006 compared to 8.2% between 1973 and 1996 (Pohl *et al.* 2010). In the USA, the group most at risk of developing adenocarcinoma is Caucasian males, with a male:female ratio of 7:1 and Caucasian:Black ratio of 4:1 (reviewed in Holmes and Vaughan 2007). Most males are diagnosed with adenocarcinoma aged between 75 and 79, with females slightly older at 80-84 (reviewed in Holmes and Vaughan 2007).

1.1.1.2 Environmental risk factors

The large increase of oesophageal adenocarcinoma cases over a relatively short period of time, as described above, suggests that environmental risk factors play an important role in the development of the disease. The strongest risk factor is gastroesophageal reflux disease (GERD) (reviewed in Reid *et al.* 2010). This is where the contents of the stomach (food, liquid or stomach acid) moves back into the oesophagus via the lower oesophageal sphincter, which normally remains closed preventing it occurring. It is thought that GERD causes damage to the oesophageal mucosa (reviewed in Hamilton and Aaltonen 2000). Individuals with GERD have a greater than four-fold risk of developing oesophageal adenocarcinoma than controls, although less than half of cancer

patients experience frequent GERD symptoms (Lagergren *et al.* 1999; Farrow *et al.* 2000; Reid *et al.* 2010). Gastroesophageal reflux disease is also believed to lead to the development of Barrett's oesophagus, a specialized intestinal metaplasia where metaplastic columnar epithelium, which is similar to the intestinal cell structure, replaces the normal oesophageal squamous epithelium. However, a recent study showed that almost half of Barrett's oesophagus cases had no GERD symptoms (Ronkainen *et al.* 2005; Zagari *et al.* 2008).

Barrett's oesophagus has itself previously been considered the main risk factor for adenocarcinoma (Hamilton and Aaltonen 2000). However, a recent review by Reid *et al.* (2010) has questioned this, as Barrett's oesophagus rarely develops into cancer and the majority (95%) of adenocarcinoma patients do not have the condition. Therefore, the relationship between adenocarcinoma, Barrett's oesophagus and GERD is still not fully understood.

A high body mass index (BMI) is also a known environmental risk factor for adenocarcinoma (reviewed by Holmes and Vaughan 2007, and Umar and Fleischer 2008). As BMI increases, the risk of developing cancer also increases, with one study reporting an odds ratio (OR) of 6.1 (95% confidence interval (CI) = 2.7 – 13.6) for individuals with a BMI of 40+ (BMI ≥ 30 is clinically obese) compared to those with a healthy BMI of 18.5 – 24.9 (Whiteman *et al.* 2008). It is suggested that interactions also occur between obesity and GERD, with individuals suffering from both conditions having an even greater risk of adenocarcinoma. Compared to people with a healthy BMI and no GERD, the following risks were observed: obese people with regular GERD symptoms, OR = 16.5 (95% CI = 8.9 – 30.6); obese people with no GERD, OR = 2.2 (95% CI = 1.1 – 4.3); and people with a healthy BMI with GERD, OR = 5.6 (95% CI = 2.8 – 11.3) (Whiteman *et al.* 2008).

Other risk factors also contribute to adenocarcinoma development but to a lesser extent than those described above. Smoking is suggested to increase

risk by 1.5 to 4-fold, with the risk remaining even 30 years after smoking cessation (Gammon *et al.* 1997; Holmes and Vaughan 2007). However, in a Swedish population, smoking was not found to be a risk factor (Lagergren *et al.* 2000). Diets with nutritional deficiencies, such as those low in fruit and vegetable intake and cereal fibre, also show evidence of association with adenocarcinoma (reviewed in Pera *et al.* 2005). Alcohol consumption is not a risk factor for adenocarcinoma, in contrast to oesophageal squamous carcinoma (reviewed in Holmes and Vaughan 2007).

The use of aspirin and other non-steroidal anti-inflammatory drugs (NSAIDS) has been shown to have a preventative role in the development of both oesophageal adenocarcinoma and Barrett's oesophagus (Anderson *et al.* 2006; Duan *et al.* 2008; Sadeghi *et al.* 2008).

1.1.1.3 Genetic susceptibility

Only a few studies, using mainly relatively small number of cases, have examined the association between genetic variants and the development of oesophageal adenocarcinoma. These have been candidate gene studies which have focused on genes involved in DNA repair, cell cycle control, apoptosis and other carcinogenic pathways, but results have been conflicting (reviewed in Cheung and Liu 2009). The largest study to date genotyped 1330 SNPs in 354 genes which were involved in 14 pathways linked to carcinogenesis (Liu *et al.* 2010). Variants in *CASP7* and *CASP9* (both involved in apoptosis) were associated with the disease, together with a SNP in *PGR* (involved in the sex hormone signalling pathway) in females only. This study contained 335 cases and 319 controls so lacked power to detect variants with a modest effect.

Several studies suggest that gene-environment interactions may affect the risk of oesophageal adenocarcinoma. For example, SNPs in angiogenic genes may interact with GERD, BMI and smoking to modify disease risk (Zhai *et al.* 2012),

and variants in apoptotic genes also show evidence of interaction with GERD and smoking (Wu *et al.* 2011.a).

1.1.2 Oesophageal squamous cell carcinoma (OSCC)

1.1.2.1 Definition and incidence

OSCC is defined as "a malignant epithelial tumour with squamous cell differentiation, microscopically characterised by keratinocyte-like cells with intercellular bridges and/or keratinisation" (Hamilton and Aaltonen 2000). It is the most the common form of oesophageal cancer, occurring predominantly in developing countries. The incidence is variable, with high risk areas identified. For example, a large number of cases are observed in the 'Asian oesophageal cancer belt' which begins in Turkey and passes through Iraq, Iran, Kazakhstan and into Northern China (Mudan and Kang 2007). In the Henan province of China, mortality rates in males are greater than 100 per 100,000 people (Hamilton and Aaltonen 2000; Mudan and Kang 2007). High risk regions are also present in South Africa, particularly the Transkei region (Somdyala *et al.* 2003; Somdyala *et al.* 2010). Western countries tend to have an age-standardized incidence rate of 1-4 per 100,000 individuals but higher rates are observed, such as in certain regions of France (~11 cases per 100,000) (IARC 2007.a; Melhado 2010). In the USA, OSCC has a higher incidence in black men than Caucasian men, contrasting to that of adenocarcinoma (Cook *et al.* 2009).

In low-risk areas, the incidence of oesophageal cancer is higher in males than females, but in high-risk regions the ratio is close to unity (Mahboubi *et al.* 1973; Parkin *et al.* 2005; Islami *et al.* 2009.a). It is thought that the higher use of alcohol and tobacco among males is responsible for the gender imbalance in the low-risk regions (Kamangar *et al.* 2009). The unity observed in high-risk regions may indicate that risk factors that have a gender bias are absent, such as smoking and alcohol consumption, and that other environmental risk factors

are involved which may be specific to these regions (Wabinga *et al.* 2000). Alternatively, genetic risk factors may be important.

1.1.2.2 Clinical: Presentation, outcome and treatments

OSCC mainly occurs in the middle third of the oesophagus (50-60%), with around 30% occurring in the lower third, and a smaller proportion (10-20%) in the upper third (Lewin and Appelman 1996). In the early stages of OSCC, the tumour is known as a superficial or early carcinoma, where the tumour is confined to the mucosa or submucosa. Advanced carcinoma occurs when tumours spread beyond the submucosa, where they are exophytic, ulcerating or infiltrative (Deere 2007).

Oesophageal cancer has a poor prognosis, with a 5-year survival rate of less than 10%, which is unchanged over the last three decades (Hendricks and Parker 2002). This is probably due to the largely asymptomatic nature of the cancer in its early stages, resulting in the majority of new cases being classified as advanced disease (stages III and IV). The most common symptoms experienced at this stage are dysphagia (difficulty swallowing), odynophagia (pain with swallowing), chest and back pain during swallowing, hoarseness, a chronic cough and weight loss. At early stages of disease, surgery and radiotherapy are potential treatments but as the disease develops into advanced stage, palliative care is generally the only option.

1.1.2.3 Environmental risk factors

World-wide, the main environmental risk factors for developing OSCC are alcohol consumption and tobacco use. In the western world, these habits are estimated to be responsible for 90% of all cases (Hamilton and Aaltonen 2000). However, this estimate is significantly lower in high-risk regions; in the Henan province of China, alcohol and tobacco are estimated to account for only 1% of cases and in South Africa this figure is 50% (Hamilton and Aaltonen 2000). This

difference in the contribution of alcohol and tobacco observed between developing countries and the western world may be due to the lower number of smokers/drinkers in developing regions or the presence of other risk factors that have a larger effect. The contribution of other risk factors to OSCC development is not fully understood and is discussed below.

1.1.2.3.1 Alcohol

In 2007, the World Health Organization's International Agency for Research on Cancer (IARC) recognised that ethanol is carcinogenic to humans and that alcohol consumption causes an increased risk of upper aerodigestive tract cancers, including oesophageal cancer (IARC 2010). The product of ethanol oxidation, acetaldehyde, is also carcinogenic to humans (IARC 2010). The exact mechanism by which ethanol and acetaldehyde cause an increased risk of OSCC is not fully understood. Ethanol is suggested to cause DNA damage and has been shown to cause sister chromatid exchange in human cells, while acetaldehyde can form DNA adducts which are mutagenic and cause chromosomal aberrations (IARC 2010). Alcohol consumption alone accounts for more than 60% of OSCC cases in several countries including USA, Japan and Taiwan (Lee *et al.* 2009). Increasing levels of alcohol consumption is correlated with a higher risk of developing OSCC (Lee *et al.* 2005).

1.1.2.3.2 Tobacco

Cigarettes contain thousands of chemicals, of which more than 60 are carcinogens, and to date, 15 have been identified as being carcinogenic to humans (IARC 2004.a; U.S. Department of Health and Human Services 2010). Following persistent exposure to these carcinogens over a number of years, DNA adducts can form in individuals leading to mutations in oncogenes or tumour-suppressor genes which affect cell growth and apoptosis pathways (reviewed in U.S. Department of Health and Human Services 2010).

The IARC has concluded that tobacco smoking does cause OSCC, which was supported by a review of published literature that found an association in 52 out of 54 studies (IARC 2004.a). This was observed in several populations, including those from China, Japan, Europe and the USA. Smokers are found to have an approximately 5-fold increased risk of developing OSCC compared to non-smokers, which increases further for heavy smokers (Blot and McLaughlin 1999). Following smoking cessation, the risk of developing OSCC is reduced, with one study observing a 60% decrease in risk after five years (Lee *et al.* 2005).

Smokers who also consume alcohol are at an even greater risk of developing OSCC, with a multiplicative interaction effect between these environmental factors (Lee *et al.* 2005). An odds ratio of 20.4 was reported for individuals who smoke and drink alcohol compared to those who abstain from both activities.

1.1.2.3.3 Diet

The effect of an individual's diet on cancer risk has produced conflicting results, perhaps because it relies on patients keeping an accurate record of food and drink consumption. A high intake of fresh fruit and vegetables has been thought to decrease the risk of cancer and other diseases, which led to the World Health Organization recommending a minimum intake of five portions of these foods a day. However, recently the protective effect of fruit and vegetable consumption against cancer risk has been shown to be minimal in a European population (Boffetta *et al.* 2010). For OSCC, results have also been inconsistent; studies in China and Uruguay have shown fruit and vegetable consumption to be either protective or to have no effect (De Stefani *et al.* 2003; Li and Yu 2003). Consumption of poultry and fish have also been associated with a decreased risk of OSCC, whereas an increased risk of disease was observed with a high intake of red, stewed and salted meat (De Stefani *et al.* 2003). Conflicting results have been found when analyzing the effect of hot drink consumption with

OSCC development (Islami *et al.* 2009.b). In an Iranian population, regularly drinking hot (>65°C) tea was associated with an increased risk of developing OSCC compared to those who drank tea at less than 65°C (Islami *et al.* 2009.c). This association was also observed in a South American population (Castellsague *et al.* 2000), but not in a Swedish population (Terry *et al.* 2001).

1.1.2.3.4 Other potential risk factors

Betal quid chewing is common in Asia where it is chewed with tobacco in most countries apart from Taiwan. Studies in the Taiwanese population have allowed the effects of betal quid chewing to be observed independently of tobacco smoking, and identified the practice as being associated with a 2-3 fold increased risk of OSCC (Lee *et al.* 2005). When combined with alcohol consumption and tobacco smoking, an odds ratio of 41.2 (95% CI = 23.6 – 72.0) was observed compared to individuals who abstain from all of these practices. For people who smoke and drink but do not chew betal quid, an odds ratio of 20.4 (95% CI = 12.7 – 32.9) was reported (Lee *et al.* 2005). The IARC concluded that using betal quid with tobacco causes OSCC (IARC 2004.b).

The presence of Helicobacter pylori (*H. pylori*) infection in the stomach has shown conflicting evidence of an association with OSCC (reviewed in Kamangar *et al.* 2009). However, three meta-analyses have found no overall association (Rokkas *et al.* 2007; Islami and Kamangar 2008; Zhuo *et al.* 2008). Similarly, association with human papillomavirus (HPV) infection has shown variable results which Kamangar *et al.* (2009) suggest may be due to a number of factors including geographic location, inadequate adjustment for other risk factors, false-positive results due to experimental error or due to chance. The IARC concludes that there is a lack of evidence to support a role for HPV in oesophageal cancer development (IARC 2007.b).

BMI has also been associated with OSCC risk, with a lower BMI associated with a higher risk of disease, which is in contrast to oesophageal adenocarcinoma (Smith *et al.* 2008).

1.1.2.4 Genetic susceptibility

A contribution of genetic variants to OSCC susceptibility is supported by the existence of individuals who are exposed to environmental risk factors but who do not develop the disease, and by others who avoid known risk factors but do develop cancer. Additionally, there is evidence that OSCC aggregates in families, with family history causing an approximate 2-4 fold increased risk of developing the disease (Chang-Claude *et al.* 1997; Hemminki and Jiang 2002; Garavello *et al.* 2005; Akbari *et al.* 2006; Wen *et al.* 2006; Gao *et al.* 2009; Wu *et al.* 2011.b). These familial cases are thought to develop earlier and have a worse prognosis than sporadic cases (Wen *et al.* 2006; Wen *et al.* 2009). However, whether a family history of disease is due to a shared environment or genetics is also an area of contention. In one Chinese population, environmental risk factors were estimated to be responsible for the majority of familial cases (Wu *et al.* 2011.b). In another Chinese population, genetic susceptibility is considered the most important factor, due to incidences of cancer in non-blood relatives not being associated with risk of OSCC (Gao *et al.* 2009). This is supported by a high-risk Iranian population, where 82% of familial cases were related by blood, with only 18% by marriage (Ghadirian 1985). Studies in monozygotic and dizygotic twins would enable heritability to be estimated, but this has not been completed for OSCC.

Genetic association studies in OSCC have predominantly been carried out in Asian populations due to the high-risk regions in this area. The majority of studies are candidate gene association studies, focusing on genes involved in alcohol metabolism, detoxification of carcinogens, DNA repair, apoptosis and cell proliferation (reviewed in Lao-Sirieix *et al.* 2010). However, the results are

not always consistent, with variations observed in different populations. This may be due to genetic risk factors and gene-environmental interactions that are specific to each population. Alternatively, studies may be underpowered to detect modest effects. OSCC candidate gene studies are discussed further in Chapter 3 of this thesis.

Several genome-wide association studies (GWAS) for OSCC have been performed in Asian populations. The first, in 2009 in a Japanese population, identified associations with *ALDH2* Glu504Lys (rs671) on chromosome 12q24 and *ADH1B* Arg48His (rs1229984) on chromosome 4q23 (Cui *et al.* 2009). This was followed by three independent studies in Chinese populations which reported association of a total of 8 single-nucleotide polymorphisms (SNPs) in 6 susceptibility loci (Abnet *et al.* 2010; Wang *et al.* 2010.a; Wu *et al.* 2011.c). One variant, *PLCE1* His1927Arg (rs2274223) on chromosome 10q23, was significantly associated in all three studies with an odds ratio of 1.34 in two studies and 1.43 in the third. This was the only SNP associated with OSCC in the study by Abnet *et al.* (2010) which used 2,115 cases and 3,302 controls, and interestingly, the authors found the same variant to be associated with gastric cancer with a similar effect size. In addition to *PLCE1* His1927Arg, Wang *et al.* (2010.a) identified an association with *C20orf54/SLC52A3* (rs13042395) on chromosome 20p13. Wu *et al.* (2011.c) found associations of six further variants with OSCC at four loci: *PDE4D* (rs10052657) on chromosome 5q12, *RUNX1* (rs2014300) on chromosome 21q22.3, a variant near *UNC5CL* (rs10484761) on chromosome 6p21.1 and three SNPs at a locus on 12q24 - *ACAD10* (rs11066015), *C12orf51* (rs2074356) and rs11066280.

Further replication studies of *PLCE1* and *C20orf54* have been carried out in Chinese and American populations. In a Chinese population (380 cases and 380 controls), only the *PLCE1* His1927Arg was shown to be associated with OSCC (GG vs. AA genotype), with an odds ratio of 1.95 (95% CI = 1.05 – 3.59) (Gu *et al.* 2012). In the US population, a small study of 52 OSCC cases and 211

controls showed an association between the *PLCE1* variant and the disease using a dominant model (OR = 0.5, 95% CI = 0.3 – 1.0; P<0.05), although the authors acknowledged that they did not correct for multiple testing (3 SNPs in 3 tumour types were tested) (Palmer *et al.* 2012). The effect is in the opposite direction to that observed in Chinese populations, and increasing the sample size in this study would help to confirm whether this is a true association.

During the preparation of this thesis, an additional study by Wu *et al.* (2012.a) was published which followed up a further 169 SNPs that showed a suggestive evidence of association with OSCC (10^{-7} < P< 10^{-4}) in their 2011 GWAS (Wu *et al.* 2011.c). Two replication stages were carried out, totalling ~8,000 of both cases and controls. A meta-analysis of the GWAS and the two replication stages produced 15 variants which were significantly associated with OSCC at the genome-wide significance threshold (P<5 x 10^{-8}). Of these, 8 were in the region of the alcohol dehydrogenase (*ADH*) gene cluster in chromosome 4q23. Additionally, a genome-wide gene-environment analysis was performed to study the effect of alcohol consumption. This found two variants, in *IGFB2* and *SLC10A2*, which were protective in non-drinkers but increased risk in drinkers. These variants were not significant when results were not stratified by alcohol consumption, suggesting that environmental factors should be taken into consideration at an early stage of the analysis.

In a population of European descent, a GWAS has been completed in upper aerodigestive tract cancers (UADT), including 437 OSCC cases (McKay *et al.* 2011). Of the 5 SNPs that were associated with all UADT cancers, SNPs at *ADH1B* (rs1229984), *ADH7* (rs1573496) and *ALDH2* (rs4767364) were associated with OSCC.

1.2 Cancer in Africa

In 2008, it was estimated that ~715,000 new cancer cases and ~542,000 cancer deaths occurred in Africa (Ferlay *et al.* 2010.b). However, cancer incidence and mortality in the continent are difficult to estimate due to the lack of cancer registries. Only 5 national and 50 local registries, covering 8% of the African population, were included in the above estimates (Ferlay *et al.* 2010.b; Jemal *et al.* 2012). Data quality in these local registries is also known to vary, with only 5 being of high enough standard for inclusion in other world-wide cancer incidence reports (Jemal *et al.* 2012). However, cancer is seen as a growing health problem in Africa, with incidences and deaths expected to almost double by 2030, with 1.34 million new cancer cases and 1.02 million cancer deaths predicted to occur (Ferlay *et al.* 2010.b). This is due to both the increase in population size over this period and the expected increase in life-expectancy resulting from a reduction in death rates from other diseases (IARC 2008).

Another important factor in the increase of new cancer cases in Africa is the changing of lifestyle habits, and as African countries develop, inhabitants with disposable incomes may adopt a more westernized lifestyle (reviewed in Jemal et al. 2012, and Lingwood *et al.* 2008). This brings its own problems, such as tobacco use, excessive alcohol consumption and obesity caused by an unhealthy diet and lack of physical activity. All of these contribute to cancer development, as well as a plethora of other health issues. Diseases resulting from these risk factors are preventable, and, therefore, will cause a burden to the health care systems in African countries which are already over-stretched due to long-term health problems such as HIV infection and malaria.

The lack of good-quality health care in Africa to diagnose and treat cancer patients may ultimately lead to premature death. The vast majority (80%) of new cancer patients in Africa are diagnosed when the disease is already at an advanced stage when palliative care is generally the only option (Jemal *et al.*

2012). This results in a low 5-year survival rate for patients which will not improve until the health care system is developed to be able to diagnose cancer earlier and to provide adequate treatment.

Currently, African governments and international funding agencies focus on communicable diseases which affect a larger number of people and will, therefore, have a greater impact, but this leaves little consideration for cancer patients. Jemal *et al.* (2012) highlight the fact that developing countries can learn from Western countries to prevent cancer becoming a major problem. For example, controlling tobacco use by banning smoking in public places, increasing taxes and educating the population regarding health issues associated with it may prevent a smoking culture from developing.

1.2.1 Oesophageal cancer in Africa

1.2.1.1 Incidence

Oesophageal cancer is ranked as the fourth and seventh most common form of cancer in African males and females, respectively (Ferlay *et al.* 2010.b). However, the incidence does vary widely across the continent. Middle and western Africa have the lowest rates in the world (~1.5 cases per 100,000 for males in both regions in 2008), whereas southern and eastern Africa observe some of the highest rates (22.3 cases per 100,000 and 14.9 cases per 100,000, respectively for males) (Ferlay *et al.* 2010.a).

In southern Africa, the disease is the third most common cancer in both males and females, with age-adjusted incidence rates in 2008 of 22.3 and 11.7 per 100,000, respectively (Ferlay *et al.* 2010.b). Certain regions of South Africa have a particularly high incidence, namely the former Transkei region in the Eastern Cape province (Somdyala *et al.* 2010). An increase in the incidence of oesophageal cancer was first reported in this area in 1957 by R.J Burrell (Burrell 1957) and, since then, the incidence has remained consistently high (Somdyala

et al. 2010). Rates within the Transkei itself also vary, with the south-western districts having a higher incidence than north-eastern regions (Somdyala *et al.* 2003). In the high-risk regions of the Transkei, age-standardised incidence rates reached 102.6 cases per 100,000 in males during 1965-1969 (IARC 2003). More recent figures for rural populations in the Eastern Cape Province, where the majority are Xhosa-speaking indigenous Black Africans, show incidence rates of 32.7 cases per 100,000 in males and 20.2 cases per 100,000 in females during 1998-2002 (Somdyala *et al.* 2010). This places osophageal cancer as the most common form of cancer in males and second in females in this region, with squamous cell carcinoma being the predominant subtype, accounting for 87.6% of oesophageal cancers (Figure 1.2) (Somdyala *et al.* 2010).

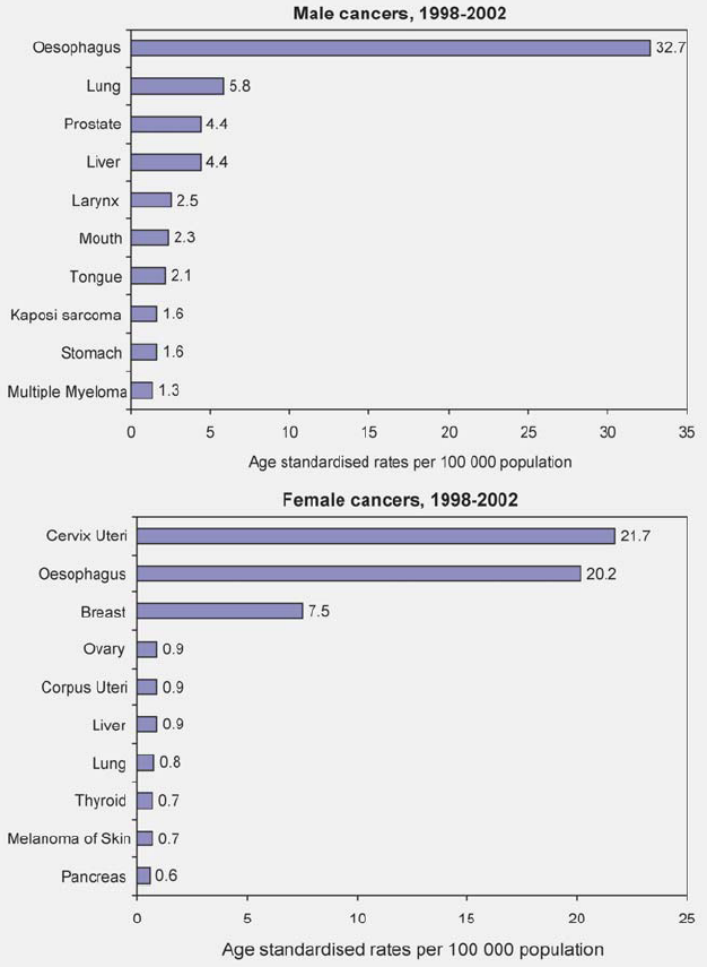


Figure 1.2: Cancer incidence in the Eastern Cape Province of South Africa
Cancer incidence in rural populations of the Eastern Cape Province of South Africa, 1998-2002 (Somdyala et al. 2010). Oesophageal cancer is the most common form of cancer in males and third in female.

1.2.1.2 Environmental risk factors

Along with recording incidences of oesophageal cancer in the Transkei, R.J. Burrell began investigating possible environmental risk factors in the 1950s. This involved visiting many villages over an area of 16,500 square miles with permission from various chiefs (Burrell 1962). Photographs taken from this period are shown in Figure 1.3, which show the typical surroundings visited and Burrell himself.

a)

FIGURE 1.—A group of Bantu typical of those surveyed. (Photo by Lister Hunter)

b)

FIGURE 2.—Dr. R. J. W. Burrell (deceased), with two of his more sophisticated Bantu friends. The man, now a public prosecutor, leans against a new, ox-drawn sledge. (Photo by W. A. Roach)

Figure 1.3: R.J Burrell investigating oesophageal cancer in South Africa
R.J Burrell visited South African villages in the 1960s to investigate risk factors for oesophageal cancer. A typical village is shown **(a),** together with R.J Burrell on a visit **(b)**. (Burrell 1969)

Incidences of oesophageal cancer together with factors such as composition of land, place of work and diet were recorded which allowed a comparison between high and low-risk areas. In 1962, Burrell proposed that factors in household gardens were responsible for the high incidences of OSCC, which would require further study (Burrell 1962). Indeed, by comparing gardens in regions with high cancer incidences to those of low incidences an association was found between OSCC and mineral deficiency in garden plants (Burrell *et al.* 1966).

In the decades following the studies conducted by Burrell, further environmental risk factors have been investigated. The *Fusarium* species of fungus, which produce mycotoxins such as Fumonisin B_1 and B_2, have been shown to contaminate maize, a staple food in the Transkei. The incidence of *F. verticillioides* in corn was found to be higher in the regions of the Transkei with a high incidence of OSCC compared to those with a lower cancer rate (Marasas *et al.* 1981). Other studies also support this, together with evidence of higher levels of mycotoxins in corn found in high-risk OSCC regions (Marasas *et al.* 1988; Sydenham *et al.* 1990; Rheeder *et al.* 1992). However, whether the fungus and the mycotoxins are causal to OSCC is unknown. The most recent IARC monograph in 2002 concluded that Fumonisin B_1 is possibly carcinogenic to humans (IARC 2002). The studies described above were all completed over 20 years ago, and it is not known whether this could still be a risk factor in the Transkei today.

Human papillomaviruses may also play a role in OSCC development in South Africa. Hale *et al.* (1989) observed HPV infection in 65% (13/20) of OSCC tumours, with Matsha *et al.* (2002) reporting rates of 46% (23/50). However, both of these studies failed to include control samples and, hence, the rate may represent a high HPV rate in the population.

The effects of alcohol consumption and tobacco smoking, the main risk factors for OSCC world-wide, have been investigated in South Africa. Results have been conflicting, particularly for alcohol consumption with either an increased risk or no effect observed, perhaps due to low sample numbers or to different forms of alcohol being consumed. Two of the most recent studies in the Black South African population show conflicting results, with Dandara *et al.* (2006) reporting no association with smoking and drinking, compared to Pacella-Normal *et al.* (2002) who reported an increased risk in smokers and drinkers. The latter study observed odds ratios of 3.8 (95% CI = 2.3-6.1) and 1.8 (1.2-2.8) for smoking and drinking, respectively. Aside from a lack of power to detect associations, a possible reason for the discrepancy observed is that there are regional differences within the country. Individuals in the study by Dandara *et al.* were from the Eastern and Western Cape, whereas the patients in the Pacella-Norman *et al.* study resided in greater Johannesburg. In the high-risk region of the Transkei in the Eastern Cape, alcohol was also not identified as a risk factor (Sammon 1992; Matsha *et al.* 2006). In these South African populations, traditional home-brewed beer was consumed which may have a different composition to westernized beer or have a lower alcohol content. This might produce a less carcinogenic effect compared to that observed in other populations where alcohol is a risk factor. Additionally, in the former Transkei region, only 16% of men regularly consumed alcohol and 31% smoked tobacco in 2002 (Somdyala *et al.* 2010), which may have led to a lack of power to detect these environmental effects. Alternatively, other environmental or genetic risk factors contribute to OSCC development in this region.

1.2.1.3 Genetic susceptibility

Several candidate gene association studies have been published which focus on OSCC in South African populations. The samples in all of these studies (including those used in this thesis) come from the same source, with the sample size increasing over time.

In 2005, the first OSCC association study in the South African population was published. This found the variant *CYP2E1*6* to be associated with OSCC in the South African populations (Li *et al.* 2005). However, this study analyzed the Black and Mixed Ancestry populations jointly. It is now believed that the populations should be analyzed separately due to our acquired knowledge of population structure which shows that the Black and Mixed Ancestry populations are genetically different. A combined analysis may lead to false-positive associations (and false-negatives) due to population stratification.

A further study reported association of the *CYP3A5*3* allele with a reduced risk of OSCC in the Mixed Ancestry population (P = 0.025) but this was not corrected for the multiple tests performed (Dandara *et al.* 2005).

Variants in alcohol metabolising genes have also been tested for association with OSCC, with the *ALDH2* rs671 variant (Glu504Lys) reported to be associated with the disease in the Black population (Li *et al.* 2008). However, this was a surprising result since other evidence suggests that this variant is specific to Asian populations (Li *et al.* 2009.a).

Two studies have investigated variants in *GST* genes for susceptibility to OSCC (Li *et al.* 2010.a; Matejcic *et al.* 2011). These are plausible candidate genes for cancer susceptibility as glutathione S-transferases (GSTs) cause the detoxification of toxic compounds, such as environmental carcinogens. The association of GST variants with susceptibility to cancer has also been studied for many years (Rebbeck 1997). The first study in the South African population identified a deletion in *GSTT1* (GSTT1*0) and the variant *GSTP1* 341 C/T (Ala114Val) as being associated with OSCC in the Black population (Li *et al.* 2010.a). This latter variant was also associated in the Mixed Ancestry population, together with a deletion in *GSTM1* (GSTM1*0). In contrast, in a more recent study by Matejcic *et al.* (2011), the 54-kb deletion in *GSTT1* was

not associated with OSCC in the Black population. This difference is thought to be due to an improved genotyping technique that allowed all three genotypes to be determined, rather than only homozygous deletions as in the earlier study, as well as an increased sample size. Also in this study, a 37-kb deletion in *GSTT2B* was significantly associated with OSCC in the Mixed Ancestry population (OR = 0.71, 95% CI = 0.57 – 0.90; P = 0.004).

Finally, variants in mismatch repair enzymes have been investigated (Vogelsang *et al.* 2012). Three variants were significantly associated with OSCC in the Mixed Ancestry population; *MSH3* rs26279, *PMS1* rs5742938 and *MLH3* rs28756991. No variants were associated with the disease in the Black population.

1.3 Population genetics

1.3.1 Modern human evolution

Modern humans are thought to have evolved in Africa about 200,000 years ago. Fossil evidence suggests that modern humans originated in East Africa but recent genetic studies indicate a southern African origin (Henn *et al.* 2011). Migration out of Africa, to Eurasia and to the Americas, is thought to have occurred within the last 40,000 - 80,000 years and 15,000 - 30,000 years, respectively, and it is estimated that only around 1,000 - 1,500 founding individuals left Africa (Campbell and Tishkoff 2010).

Polymorphisms that arose in African populations prior to migration out of the continent account for ~90% of variation seen in all human populations today (McClellan and King 2010). The migration of a relatively small number of individuals out of Africa led to a reduction in genetic variation in subsequent generations of the migrants, known as a population bottleneck. Variation that has arisen since then is thought to be due to rapid population growth caused by the development of agriculture and urbanisation (McClellan and King 2010),

which will be influenced by factors specific to a population such as the environment and selection pressures. Due to the population bottleneck that occurred, African populations are the most genetically diverse in the world (Tishkoff and Williams 2002).

1.3.2 Genetic structure of African populations

The first large-scale study into the variation in African genomes was the HapMap project which started in 2002 and aimed to identify genetic polymorphisms in multiple human populations of European, Asian and African ancestry, including the Yoruba from Ibadan, Nigeria (YRI) (www.hapmap.org). On a similar principle, the 1000 Genomes project aims to whole-genome sequence 2,500 individuals from 27 global populations, including five from Africa (www.1000genomes.org).

Numerous other studies have also focused on determining the structure of African populations to reveal insights into the origin of modern humans, evolutionary history and adaptation of populations (Tishkoff *et al.* 2009; de Wit *et al.* 2010; Quintana-Murci *et al.* 2010; Henn *et al.* 2011; Lachance *et al.* 2012; Pickrell *et al.* 2012; Schlebusch *et al.* 2012). In one study, the hunter-gatherer populations (including the Hadza and Sandawe of Tanzania and the ≠Khomani Bushmen of South Africa) were found to have the greatest amount of genetic diversity in the world and have the lowest linkage disequilibrium levels across 27 African populations (Henn *et al.* 2011). From this work, Henn and colleagues were able to conclude that modern humans were likely to have originated in southern Africa.

In 2010, the first whole genome sequences of African individuals were published of a Khoisan hunter gatherer and a Bantu individual (Archbishop Desmond Tutu), together with exome sequences from a further three hunter-gatherers, all from southern Africa (Schuster *et al.* 2010). In total, 1.3 million novel variants

were discovered, and remarkably, the authors found that the hunter-gatherers were more genetically distinct from each other than a European and an Asian individual.

More recently, Lachance *et al.* (2012) have sequenced the whole genomes of 5 individuals from three hunter-gatherer populations; one from Cameroon (Pygmies) and two from Tanzania (Khoisan-speaking Hadza and Sandawe). As well as identifying novel variants and determining shared ancestry both between the hunter-gatherer populations and within other African populations, this approach enabled the authors to identify loci that were specific to traits, such as height in the Pygmy population.

1.3.3 Population history and genetic structure of the South African Xhosa and Mixed Ancestry populations

The hunter-gatherers populations, known as the San (or Bushmen) and Khoikhoi, or collectively as the Khoisan, were the first inhabitants of what is nowadays known as South Africa. Bantu-speaking groups originated from Cameroon/Nigeria and began to migrate around 1500 BC, spreading both south and to the east (Johnson 2004; Bryc *et al.* 2010). It is the latter group that eventually reached South Africa in the third century. Five Bantu-speaking groups were known to reside in South Africa, all of which relied on agriculture and cattle herding (Johnson 2004). One of these groups, the Nguni, is the origin of the Xhosa and Zulu populations, and migration continued through the east of South Africa, with the Xhosa settling in what is now the Eastern Cape Province. The Xhosa and Zulu languages are influenced by the click-sounding Khoi languages, showing that the groups would have interacted. Indeed, four studies have analyzed the genetic structure of the Xhosa population, and, although the number of other African populations in each study is variable, results show that the Xhosa contain Khoisan ancestry (Tishkoff *et al.* 2009; Bryc *et al.* 2010; Patterson *et al.* 2010; Schuster *et al.* 2010). Two of the studies also suggest

relatedness with the Yoruban population (Patterson *et al.* 2010; Schuster *et al.* 2010).

In 1652, non-African populations began to arrive in the area now known as Cape Town due to the opening of a fuelling station by the Dutch East India Company to allow their ships to be re-supplied on route to East Asia (reviewed in de Wit *et al.* 2010). This resulted in the settlement of Europeans which was followed in 1658 by the importation of slaves from India, Indonesia, Madagascar and other African countries. This began the formation of the South African Mixed Ancestry population, described historically as the 'Coloured' population. With the settlers being mainly male, mixed marriages between white Europeans and local Khoi females were initially encouraged (Johnson 2004). Genetic studies show 60% of the maternal contribution to the Mixed Ancestry population is from the Khoisan, with a negligible contribution from European females (Quintana-Murci *et al.* 2010). Slaves were also encouraged to have children to provide more labour, with the white slave owners often fathering the child (Johnson 2004). Race-based restrictions first came into force in the 1700s and in the twentieth century this extended to the banning of inter-racial marriages and forcing population groups to reside in particular areas. These restrictions over several centuries established the Mixed Ancestry population. Restrictions were not alleviated until the post-apartheid era in 1994.

Four studies have analysed the genetic structure of the Mixed Ancestry population, and these support the historical evidence of the formation of the population (Tishkoff *et al.* 2009; de Wit *et al.* 2010; Patterson *et al.* 2010; Quintana-Murci *et al.* 2010). The studies do differ slightly in the estimates of the amount of each ancestral component, which may be due to sampling of the Mixed Ancestry population from different regions or as a result of using different world-wide populations in the analysis. De Wit *et al.* (2010) estimated the ancestral components as Khoisan (32-43%), Bantu-speaking Africans (20-36%), European (21-28%) and Asian (9-11%), whereas Tishkoff *et al.* (2009)

estimated approximately equal levels of Khoisan, Black African, European and Indian ancestry with a lower level of East Asian ancestry. Patterson *et al.* (2010) did not quantify the contribution of each component but reported genetic contributions from at least the Xhosa, Europeans, South Asians and Indonesians. Finally, Quintana-Murci *et al.* (2010) determined five ancestral populations; Khoisan, Bantu, European, Indian and Southeast Asian.

The most recent figures (2011 census) show that South Africa has a population of over 50 million people and consists of different population groups as follows: 79.2% Black Africans, 8.9% Mixed Ancestry, 8.9% white and 2.5% Indian/Asian and 0.5% other (http://www.statssa.gov.za/Census2011/Products/Census_2011_Census_in_brief.pdf). IsiZulu is the most frequently spoken first language (22.7%), followed by isiXhosa (16%), Afrikaans (13.5%) and English (9.6%). The Xhosa mainly reside in the Eastern Cape, with over 5 million of the 8 million individuals who speak Xhosa as a first language living here, making up 78.8% of the Eastern Cape population. The majority of those who speak Afrikaans as a first language mainly reside in the Western Cape, making up ~50% of the population in this region.

1.3.4 Genetic association studies in African populations

The genetic structure of African populations is characterised by high levels of haplotype diversity and low levels of linkage disequilibrium (LD) (Teo *et al.* 2010). This is a disadvantage in the identification of disease associations in GWAS since these arrays are currently designed based on European ancestry. GWAS arrays may exclude SNPs in complete LD with another variant as no additional information is gained from their inclusion. Hence, causal variants may not be present on the array, but their effect will still be detected through the analysis of a SNP in a high level of LD with it (known as tagging SNPs). In African populations with a lower level of LD, tagging SNPs for causal variants

may not exist and so no association will be detected. Alternatively, a tagging SNP may be in moderate LD with the causal variant in African populations but due to the weaker effect that would be observed, it may fail to reach the genome-wide significance threshold to account for multiple testing (Teo *et al.* 2010). Indeed it is noted that it is harder to achieve genome-wide significance in African populations compared to European and Asian populations due to the lower level of LD (Jallow *et al.* 2009). To successfully use a GWAS in African populations, a larger number of SNPs are likely to be needed to obtain significant coverage of their genomes. SNP arrays should also be designed based on sequence data from the relevant populations, which for the majority of African populations is lacking.

However, there is an advantage to performing association studies in African populations as they provide an opportunity to identify causal variants in disease susceptibility. Current genome-wide association studies in non-African populations are often unable to distinguish the causal variant from tagging SNPs. In these populations, the high level of LD may extend over a large region, meaning that tagging SNPs may not even be located in the gene containing the causal variant. If the regions that show disease associations in non-African populations are fine-mapped in African populations, then causal polymorphisms could potentially be identified. This is summarized in Figure 1.4 where the causal variant is in high LD with a nearby variant in non-African populations, potentially leading to this tagging variant being associated with disease.

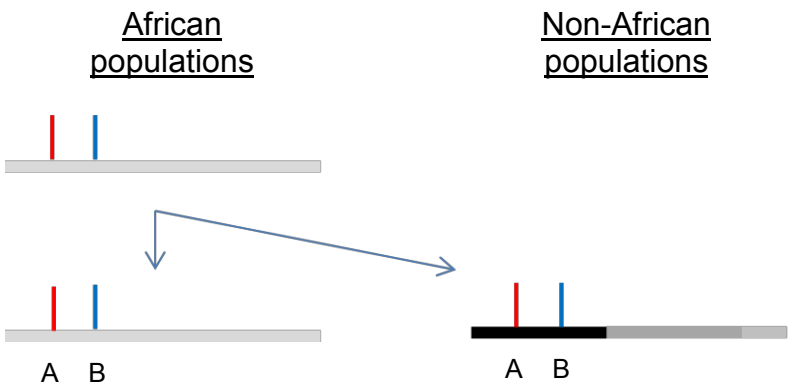


Figure 1.4: Linkage disequilibrium in African populations
Due to the population bottleneck that occurred following migration out of Africa, non-African populations have a higher level of linkage disequilibrium (LD) than African populations (as shown by the shade of the bars: the darker the colour, the higher the level of LD). In African populations, a disease causal SNP (A) and a nearby SNP (B) are independent of each other. In non-African populations, the high level of LD between these variants means SNP B 'tags' the causal SNP and hence will also be associated with disease status. If this region is fine-mapped in the African populations, association studies could potentially identify the causal rather than the tagging variant.

Promisingly, a proof-of-concept study, where the causal variant (rs334 at the haemoglobin S locus) was already known, has shown that lack of significant associations in an African GWAS can lead to the identification of the causal variant through other approaches (Jallow *et al.* 2009). This study attempted to identify the causal variant for protection against malaria in individuals from The Gambia, West Africa. A GWAS, which did not contain rs334, did not identify any variants that reached genome-wide significance, with the strongest association having $P = 3.9 \times 10^{-7}$. This region, which was also the region known to contain the causal variant, was subsequently fine-mapped in a subset of samples and the information used to impute the genotypes for an expanded set. This resulted in much stronger association signals, with the strongest association being $P= 4.5 \times 10^{-14}$, which was the causal rs334 SNP. This study highlights the fact that current GWAS chips designed for non-African populations are, alone, not suitable for association studies in African populations. Indeed Jallow *et al.* also note that genotyping arrays based on the HapMap YRI populations may also not be adequate. With our knowledge of the vast degree of genetic variation that is present in African populations, genotyping arrays specific to the population of

interest will be needed. Alternatively, the availability of reference panels for specific populations would allow imputation.

An alternative approach to identify disease associations is emerging. Pasaniuc *et al.* (2012) have shown that low-coverage (0.1-0.5x) whole exome sequencing data, together with imputation to reference panels, can be used to perform a genome-wide association study, producing similar results to using a standard GWAS array. This will enable all detected variants to be tested for disease association, which would be beneficial in African population studies.

Once the best approach for large-scale hypothesis-free association studies in African populations has been determined, it will be important to identify susceptibility variants for diseases that are prominent in these countries themselves (reviewed in Campbell and Tishkoff 2008). To date, approximately 14 genome-wide association studies have been completed in African populations (excluding African-American) with varying sample numbers (Hindorff *et al.*; accessed 04/04/2013), including analysis of the susceptibility to HIV, tuberculosis and malaria (Jallow *et al.* 2009; Lingappa *et al.* 2011; Thye *et al.* 2012; Timmann *et al.* 2012). No studies have yet examined susceptibility to cancer.

1.4 Genetic susceptibility to cancer

1.4.1 Models of susceptibility

In 1971, Alfred Knudson proposed that cancer can be caused by the acquisition of two mutations in the same gene, later known as the 'two-hit hypothesis' (Knudson 1971). Here, Knudson studied retinoblastoma in patients with familial and sporadic cancers and found that tumours in familial cases were mostly biliateral and multi-focal, whereas sporadic tumours were unilateral and unifocal. Statistical analysis of the age distribution of bilateral and unilateral cases

indicated that they matched a “one-hit” curve and “two-hit” curve, respectively. He proposed that patients with the familial dominant form of the cancer inherited one germline mutation and acquired an additional somatic mutation, whereas in sporadic forms of the disease, both mutations occurred somatically. Later, experimental studies confirmed the model, showing that germline and/or somatic mutations result in two non-functional copies of *RB1*, a tumour suppressor gene (Friend *et al.* 1986). This two-hit model is relevant to other dominant forms of familial cancer which affect tumour suppressor genes, including breast cancer caused by mutations in *BRCA1* and *BRCA2 (Smith et al. 1992; Gudmundsson et al. 1995)*. In most sporadic forms of cancer, more than one gene is required to be mutated in order for the disease to develop. The two-hit model may still be applicable to result in the loss of function of a tumour suppressor gene, but other mutated genes will also contribute to tumourigenesis (reviewed in Berger *et al.* 2011).

In addition to the two-hit model, other mechanisms of action of tumour suppressor genes have also been suggested which have been extensively reviewed by Berger *et al.* (2011) and is summarized below (see also Figure 1.5). One of these mechanisms is haploinsufficiency, whereby the normal function of one copy of a gene is lost through mutation or deletion. The effect on disease development may be due to the 50% reduction in gene expression or protein activity, or due to the formation of a mutant protein that inhibits the function of the normal protein (Berger *et al.* 2011). Haploinsufficiency together with the two-hit hypothesis, represent discrete models of loss of tumour suppression, whereby either one or two copies of the gene are non-functional. The alternative to this is a continuum model of tumour suppression. This suggests that it is the change in gene expression or protein expression that is important for tumour development, which may vary between 0-100%, with a greater loss causing a higher chance of malignancy. For some tumour suppressor genes, a ‘fail-safe mechanism’ may be activated if complete loss of the gene occurs, whereby other mechanisms are activated to prevent the consequences of complete loss

of the gene. For example, if complete loss of *Pten* occurs, *p53* causes *Pten*-loss-induced cellular senescence (PICS). This continuum model may have important implications for cancer susceptibility as changes in gene expression may be caused by germline polymorphisms in the promoter or regulatory regions adjacent to, or at some distance from the relevant gene (Berger *et al.* 2011).

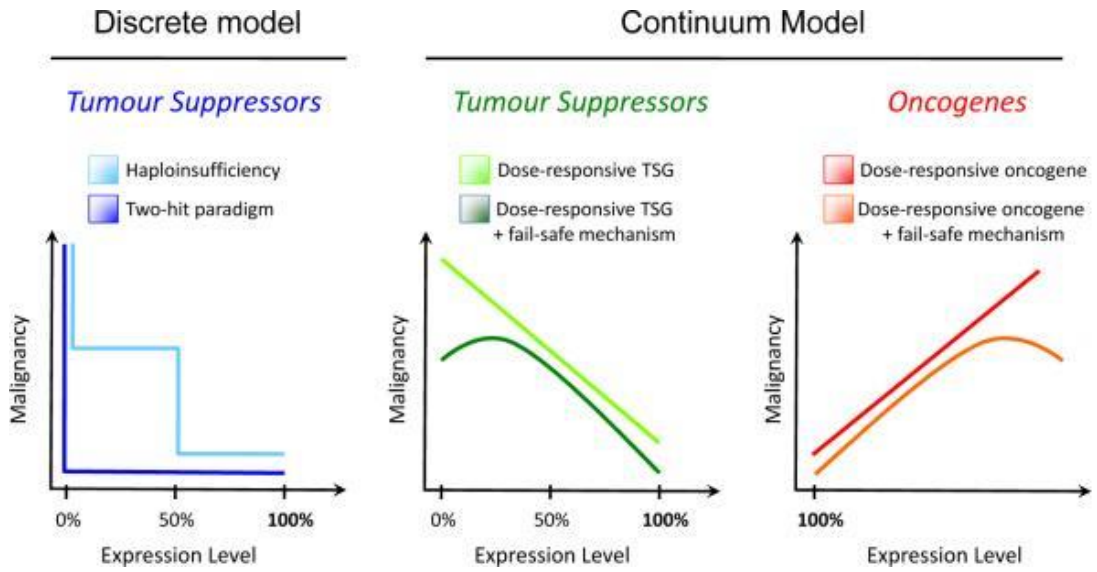


Figure 1.5: Models for tumour suppressor genes
Discrete models of tumour suppressor genes include the two-hit paradigm and haploinsufficiency, which result in one or two non-functional copies of the gene. An alternative model is a continuum, whereby tumourigenesis is correlated with a reduction in gene expression (or protein activity) levels. 'Fail-safe mechanisms' may also be activated which attempt to halt malignancy by causing cellular senescence. Oncogenes may also comply with this model. (Berger *et al.* 2011)

It is thought that cancer susceptibility in the general population is affected by the co-inheritance of multiple risk variants, together with the presence of environmental risk factors. These germline polymorphisms have a smaller effect than the highly penetrant alleles described previously for dominant familial cancer, and include rare moderate-risk alleles (minor allele frequency (MAF) <2%; OR >2.0) and common low-risk alleles (MAF >10%; OR <1.5) (reviewed in Fletcher and Houlston 2010). The types of risk alleles that have been identified in breast cancer are shown in Figure 1.6. Common low-risk variants have proved hard to detect, requiring extremely large sample numbers. In addition,

methods used to detect these variants may not identify the causal variant, as discussed in the following section.

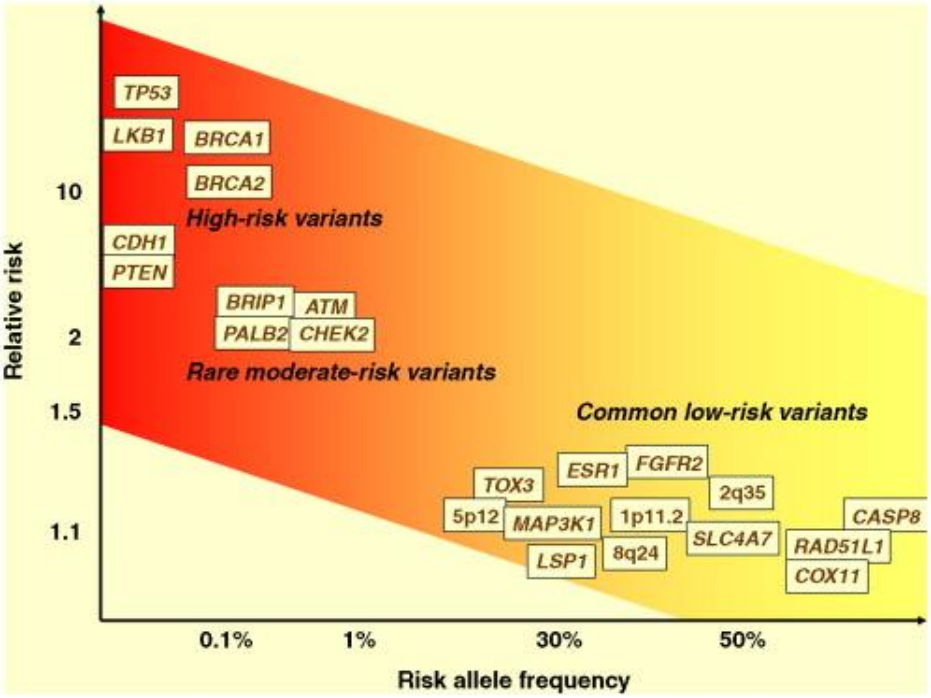


Figure 1.6: Types of susceptibility loci
Types of breast cancer susceptibility loci based on the relative risk and the risk allele frequency. (Varghese and Easton 2010)

1.4.2 Detection of susceptibility loci using association studies

1.4.2.1 Candidate gene association studies

Candidate gene studies make use of a prior knowledge of a genes function or disease associations to select genes of potential interest for a genetic association study. For example, genes involved in apoptosis or DNA repair would be good candidates for cancer studies as mis-regulation of these pathways could lead to uncontrolled growth or genomic instability in cells. Using cases and controls from the same population, the genotype and allele frequencies of a selected variant can be determined and analysed to test whether there is a significant difference between the two groups. This has led to the identification of several disease susceptibility loci. However, the number of

genes that can be tested for association has been limited by the time-consuming process of testing one gene at a time. Also, the prior probability of selecting a disease-associated gene is low, given the total number of genes in the human genome.

1.4.2.2 Genome-wide association studies

Genome-wide association studies (GWAS) are a hypothesis-free approach, enabling case-control association tests of variants across the whole genome. The number of SNPs that can be genotyped in a single experiment has increased over the years, with current platforms, such as the Illumina HumanOmni5-Quad, able to genotype 4.3 million markers per sample. Due to the large number of tests carried out, a high level of significance is required to declare a disease association, with the consensus significance threshold being $P<5 \times 10^{-8}$, based on the Bonferroni correction. In order to have high power to detect common variants which contribute a small effect, several thousand cases and controls are needed. If variants are rare or low frequency, then tens of thousands of samples will be required, which is beyond the scope of many disease studies in terms of project funding and acquiring sufficient sample sizes. Typically, a GWAS is carried out in an initial set of cases and controls and the most promising hits are then selected for replication in an independent set of samples. A meta-analysis can be performed to generate a combined p-value for association.

The current GWAS genotyping platforms were designed for populations of European descent, with the SNPs selected based on frequencies and LD in this population. If the platform is not optimized for use in other populations, true associations may fail to be detected. For example, as discussed previously (see section 1.3.4), African populations are the most genetically diverse in the world and have a lower level of LD. A variant found to be associated with a disease in a European population GWAS may be tagging the causal SNP which was not

itself genotyped. The same tagging SNP would not be associated with the disease in the African population if it was not in strong LD with the causal SNP. Therefore, population-specific GWAS genotyping platforms are needed, which will initially require the variants and haplotypes in these populations to be identified. The HapMap and 1000 Genome Projects are contributing to our knowledge in this area (www.hapmap.org; www.1000genomes.org).

According to the Catalog of Published Genome-Wide Association Studies published by the National Institute of Health, a total of 1,554 studies have identified 8,972 variants associated with diseases/traits (Hindorff *et al.*; accessed 04/04/2013).

1.4.2.3 Customized genotyping arrays

Customized genotyping arrays are also available to enable customers to select their own SNPs of interest to be included on the chip. For example, the Illumina "Infinium iSelect HD Custom Genotyping BeadChips" allows 3,000 - 1,000,000 SNPs to be selected. Arrays which have been designed in this manner and which focus on specific disease areas are the Illumina Immunochip and the Metabochip. The Immunochip was developed by a consortium of groups researching immune-related diseases, including Crohn's disease, rheumatoid arthritis and psoriasis (Cortes and Brown 2011; Trynka *et al.* 2011). The Metabochip is designed for the investigation of 23 traits focusing on metabolic, cardiovascular and anthropometric traits (Voight *et al.* 2012). Both of these genotyping platforms contain ~200,000 variants which attempt to replicate known regions of associations and those which have previously failed to meet the genome-wide significance threshold of $P<1 \times 10^{-8}$. In addition, they are designed to allow the fine-mapping of known association regions. Also included were SNPs of particular interest to any of the groups involved in the chip design.

Several disease areas have now published their results of Immunochip and Metabochip studies including psoriasis, rheumatoid arthritis, celiac disease, inflammatory bowel disease and coronary heart disease (Trynka *et al.* 2011; Deloukas *et al.* 2012; Eyre *et al.* 2012; Jostins *et al.* 2012; Tsoi *et al.* 2012). These have genotyped extremely large numbers of samples, for example in inflammatory bowel disease, 26,000 cases and 16,000 controls were used (Jostins *et al.* 2012). All studies have identified many additional variants that are associated with disease susceptibility.

1.4.2.4 Missing heritability

Despite using these large-scale genotyping platforms, the cumulative effect of susceptibility variants identified through association studies does not always match the estimated heritability of the disease calculated from twin studies. This has led to the 'missing heritability' problem which involves all complex diseases. For example, in a meta-analysis studying the susceptibility to psoriasis using >10,000 cases and >22,000 controls, the 39 independent associated variants account for 22% of the estimated heritability (Tsoi *et al.* 2012).

This missing heritability has several potential sources (reviewed in Manolio *et al.* 2009). Firstly, studies may still be under-powered to detect common variants with small effect sizes and rare variants with moderate or small effect sizes. In addition, power may be lost due to genotyping tagging variants instead of causal variants. To enable more variants to be tested for association with disease, genotypes for additional variants can be imputed using reference panels (e.g. from 1000 Genomes project) from the same population. If reference panels are not available for a population, whole-exome or whole-genome sequencing may be needed, either in the study samples themselves, or in a set of individuals to form a new reference panel where imputation can then be performed in the study samples. Inclusion of known population-specific variants on GWAS arrays

may provide additional or stronger disease associations in these populations, if imputation is not possible.

In addition to single-nucleotide polymorphisms, other factors may also be responsible for the missing heritability. These include structural variants, such as insertions, deletions, inversions and translocations, which are currently difficult to genotype accurately in large-scale studies; a Wellcome Trust Case Control Consortium found very limited evidence of the involvement of structural variants in 8 common diseases (Craddock *et al.* 2010). Other factors which might contribute to disease susceptibility are epigenetic changes, and both gene-gene and gene-environment interactions.

Alternatively, associations may be diluted by the presence of a wide range of phenotypes within the disease classification. Sub-set analysis, such as young-onset disease or variations in clinical presentation, may identify further susceptibility loci specific in these groups. However, this will result in reduced sample sizes and potentially a loss of power.

1.5 Somatic mutations in cancer

1.5.1 Cell division, errors and repair

Cell division occurs during development and throughout life to replace existing cells. Errors may occur in DNA replication due to DNA damage caused by both internal and external factors. Internal factors include the production of reactive oxygen species as a result of normal cellular metabolism. These can oxidise bases leading to mis-pairing during replication, or induce single- and double-strand breaks (De Bont and van Larebeke 2004). External sources include environmental factors known to increase the risk of cancer, such as tobacco smoke and ultraviolet radiation. Damage initiated by tobacco smoke is due to the formation of DNA adducts as a result of detoxification of the carcinogenic

compound by enzymes such as cytochrome P450. These DNA adducts bind DNA and can cause nucleotides to be mis-read during replication, resulting in mutations. Ultraviolet radiation is known to cause the formation of pyrimidine dimers which distorts the shape of the DNA helix, affecting both transcription and replication.

The majority of DNA damage and replication errors that occur will be repaired by the cell's DNA repair mechanisms, such as nucleotide excision repair. The errors which escape this process will be retained, with daughter cells inheriting them, producing germline or somatic mutations depending on the cell type in which they occur.

1.5.2 Somatic mutations

If a somatic mutation occurs in a gene or region which confers a growth advantage to the cell, this can initiate the process of cancer development. These mutations occur in genes that play a key role in pathways that are usually highly regulated, such as cell growth or apoptosis, and are referred to as driver mutations. Driver mutations occur in oncogenes and tumour suppressor genes. The former are genes are usually involved in the promotion of cell growth, whereby harmful mutations cause them to continually promote growth. Mutations in oncogenes are activating mutations which act in a dominant manner, requiring only one mutation to lead to a cellular effect. Tumour suppressor genes normally function to suppress cell growth or act as caretakers of genomic stability. Mutations occur in both copies of the gene in the same cell, hence acting in a recessive manner, and lead to inactivation of the protein. Over 80% of known driver mutations are in oncogenes (Stratton 2011).

Somatic mutations are randomly dispersed throughout the genome, and many will occur in regions which have no effect on protein function. If these occur in a

cell with a driver mutation, they will be co-selected in the clonal evolution of the tumour, and are referred to as passenger mutations.

The number of driver mutations present in tumours is known to vary in different cancer types (reviewed in Stratton 2011). For example, medullablastomas harbour only a few mutations, whereas lung cancers have many more which is thought to be due to the continual bombardment of the genome by carcinogenic compounds in tobacco smoke. As cancers evolve they will acquire more driver mutations in order for them to continue to survive and expand. This is a multi-step process which results in a neoplastic tumour that has the six hallmarks of cancer; “self-sufficiency in growth signals, insensitivity to growth-inhibitory signals, evasion of programmed cell death, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis” (Hanahan and Weinberg 2000). Recently, “reprogramming of energy metabolism and evading immune destruction” have also been added to this list (Hanahan and Weinberg 2011). To achieve all of these hallmarks, many processes have to be mis-regulated, which is perhaps why so many genes have been found to be mutated in cancer.

To date, 487 genes have been reported to harbour mutations which have been implicated in cancer, of which 84% are somatic variants, 8% are germline variants, and 8% contain both types (Cancer Gene Census; accessed 17/12/12; www.sanger.ac.uk/genetics/CGP/Census/). Many more may exist and the challenge is to identify the driver mutations and to distinguish them from the passenger mutations. Driver mutations are likely to cause significant changes to the structure or expression of a gene and be found in a significant proportion of tumours of a particular type. Therefore, genes that are frequently mutated are likely to be essential in the development of that particular cancer.

1.5.3 Why do we want to identify somatic mutations?

The identification of driver somatic mutations could provide enormous amounts of information enabling us to understand how cancer develops, and to explain the variation occurring between the same type of cancer in different patients, enabling re-classification based on mutations present. For example, one of the most thorough studies to date has been in breast cancer where the genomes and transcriptomes of ~2000 tumours were analysed (Curtis *et al.* 2012). By combining genetic copy number alternations and expression data, samples were suggested to cluster into 10 subtypes. This may lead to clinical benefits by determining the type of treatment that is most likely to be successful for each subtype.

Genetic mutations have been shown to affect treatment success. For example, colorectal cancer patients with mutations in *KRAS* do not to respond to cetuximab therapy and have a worse prognosis than those without *KRAS* mutations (Lievre *et al.* 2006). In addition, the identification of driver mutations can also provide a new therapeutic target for drug development. For example, targeting the known *BRAF* V600E mutation in melanoma patients led to tumour regression in the majority of treated patients (Flaherty *et al.* 2010).

1.5.4 Methods to detect somatic mutations

The principle of detecting somatic mutations involves comparing tumour DNA to normal DNA (either blood or normal non-cancerous tissue) from the same patient to detect changes that have arisen in the tumour. The first mutations to be observed in cancer were large chromosomal translocations that could be observed under a microscope. Following this, copy number changes could be identified using comparative genome hybridization (CGH) or, more recently, array CGH. In addition, SNP genotyping arrays can detect copy number alterations together with regions of loss-of-heterozygosity (LOH). The ability to

study small insertions/deletions and single nucleotide mutations has only been possible since the advent of DNA sequencing. Sanger sequencing enabled the analysis of candidate cancer genes, and more recently, whole-exome sequencing has allowed an unbiased approach to examine all somatic mutations occurring within the coding regions of the genome. Whole-genome sequencing of cancer patients has also been successful in identifying somatic mutations, but this approach requires additional resources, which is discussed below.

1.5.4.1 Whole-exome and whole-genome sequencing

A major advance in discovering somatic mutations has been the development of next-generation sequencing, where whole-exomes or whole-genomes can be sequenced in a relatively short time period. Whole-exome sequencing has been the favoured method to date, with the sequencing only of ~1% of the genome (~30Mb) providing numerous advantages. Firstly, sequencing the exome enriches for driver mutations by focusing on the coding regions of genes, and is quicker and cheaper than sequencing the whole genome. In addition, whole-genome sequencing produces a very large amount of data which requires more computational storage space and experienced bioinformaticians to identify the potential driver mutations amongst the thousands of somatic mutations that are likely to be present. However, whole-genome sequencing can also detect large-scale genomic changes, such as deletions and translocations, and mutations in regulatory regions outside of the exome.

The whole-genome/exome approach is not always successful in identifying driver mutations. Reasons for this include the contamination of tumour tissue with surrounding normal tissue. If this occurs, then the mutation will be present in a low percentage of the sequencing reads, which may be below the threshold for calling a variant. The threshold could be reduced, but this risks an increase in false positive mutation calls.

Another reason for the failure to identify novel driver mutations is that there tends to be a selective follow-up of mutations in genes which have already been identified as a cancer gene. Therefore, mutations in genes of unknown function or with no known link to cancer will be treated with caution and are less likely to be included in the next stage of analysis. In order to identify them as driver mutations, a large number of cancers will need to be sequenced to determine if the genes are mutated in multiple patients. To determine the effect of a variant, functional validation may be needed to assess whether processes such as growth or apoptosis are affected.

If few or no driver mutations are identified by exome sequencing, then whole-genome sequencing may unveil other classes of mutations. However, some tumour types may have very few mutations.

1.5.5 Large-scale cancer sequencing projects

Several large-scale projects exist to explore the molecular basis of cancer including The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/) and the International Cancer Genome Consortium (ICGC) (http://icgc.org/). TCGA is a comprehensive effort to sequence and identify somatic changes in more than 20 types of cancer, and the ICGC aims to identify not only the genomic changes in 50 cancer types but also the transcriptomic and epigenetic changes.

The large amount of data that is being produced from whole-genome/exome sequencing projects has led to the development of several databases which compile the results of published studies for easy access. One of these, The Catalogue of Somatic Mutations in Cancer (COSMIC) (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/), contains all the somatic mutations that have been identified in different types of cancer.

Additionally, The Cancer Gene Census (www.sanger.ac.uk/genetics/CGP/Census/) provides a list of genes that have been identified as cancer genes, together with the class of mutation and the cancer type.

1.5.6 Somatic mutations in oesophageal cancer

According to the COSMIC database (accessed 04/04/2012), *TP53* is the most frequently mutated gene in OSCC, with 50% of tumours (933/1884) harbouring mutations (Figure 1.7). This is followed by *NOTCH1* with 43% (16/37) of tumours mutated. One whole-exome sequencing study has been performed in OSCC, with 12 matched normal and tumour samples sequenced (Agrawal *et al.* 2012). After follow-up of frequently mutated genes in an additional 41 pairs of samples, Agrawal *et al.* identified the following mutation rates: *TP53* (62% of tumours), *NOTCH1* (21%), *NOTCH2* (6%), *NOTCH3 (8%)* and *FBXW7 (6%).* No OSCC samples have been whole-genome sequenced as yet, which may provide further insights into the complete mutation profile of this form of cancer. Somatic mutations in OSCC are further discussed in Chapter 6 (section 6.1).

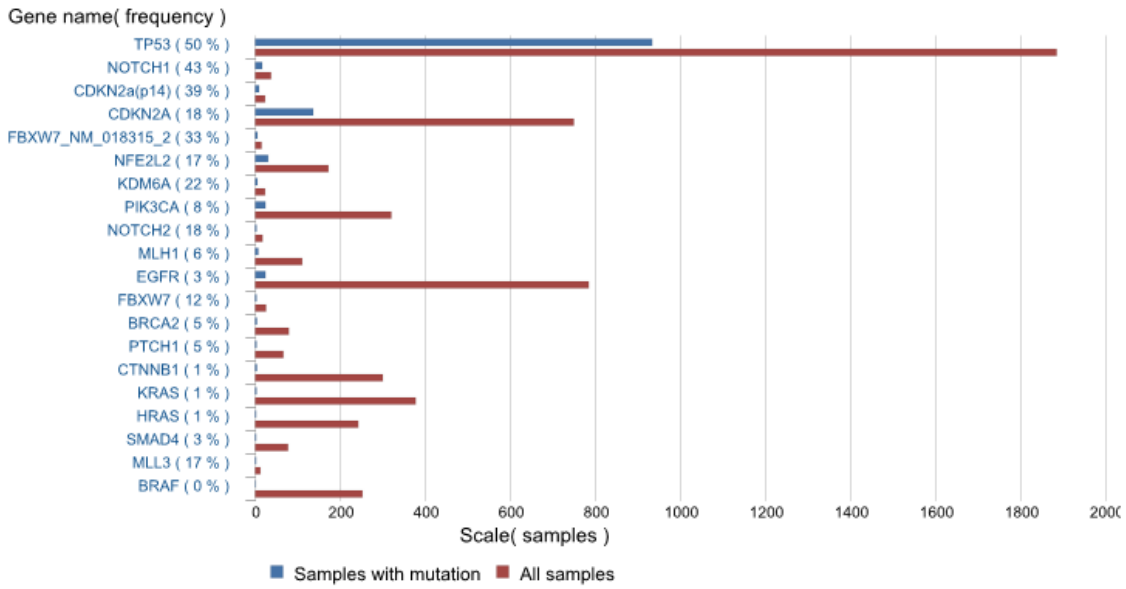


Figure 1.7: Somatic mutations in OSCC
COSMIC data for the top 20 genes mutated in 1157 OSCC samples, only including genes that are present in the Cancer Gene Census.
(http://cancer.sanger.ac.uk/cosmic/browse/tissue?sn=oesophagus)

1.6 Aims

The aim of this project is to investigate the genetics of oesophageal squamous cell carcinoma (OSCC) in the Black and Mixed Ancestry populations of South Africa. Firstly, the genetic susceptibility to OSCC will be studied by performing case-control association studies of genes and loci with strong evidence of association with OSCC in other populations, and by using the Illumina Immunochip genotyping array. The latter approach will also enable the population structure of the samples to be investigated. Secondly, somatic mutations will be identified by sequencing the whole-exome of OSCC blood-tumour pairs.

2 Methods

2.1 Materials

2.1.1 Reagents

Absolute qPCR ROX mix	ABgene
SureSelect Reagent Kit	Agilent
Herculase II Fusion DNA Polymerase	Agilent
SureSelectXT Human All Exon V4 capture Library	Agilent
BigDye® Terminator v3.1 Cycle Sequencing Kit	Applied Biosystems
Genescan™ 500 Liz™ Size Standard	Applied Biosystems
Hi-Di Formamide	Applied Biosystems
Sequencing buffer	Applied Biosystems
TaqMan SNP assays	Applied Biosystems
Agencourt AMPure XP kit	Beckman Coulter Genomics
ExoSAP-IT	GE Healthcare
TruSeq Cluster Generation Kit	Illumina
TruSeq Sequening Kit	Illumina
dNTPs	Invitrogen
Qubit® dsDNA BR (Broad-Range) Assay	Invitrogen
Qubit® dsDNA HS (High-Sensitivity) Assay	Invitrogen
Kaspar v4 High ROX mix	KBioscience
Kasp by design assay	KBioscience
Dynabeads MyOne Streptavidin T1	Life Technologies
2X PCR mastermix	Promega
GoTaq® Flexi DNA polymerase	Promega
QIAamp DNA Mini Kit	Qiagen
Orange G	Sigma-Aldrich
Reaction buffer	Thermo Scientific
Magnesium Chloride	Thermo Scientific
100 bp DNA Ladder	Web Scientific

Dimethyl sulphoxide (DMSO)	VWR

2.1.2 Solutions

- 10 x Tris/Borate/EDTA (TBE):
  - Tris-HCL 108g
  - Boric Acid 55.6g
  - EDTA 9.3g
  - Made up to 1 litre with Milli-Q water
- DNA /RNA Precipitation Solution:
  - 75µl NaOAc
  - 1562.5µl 95% Ethanol
  - 362.5µl Milli-Q water
- Orange G Loading Buffer:
  - Glycerol 30%
  - Orange G 0.2% w/v
- 10x Phosphate Buffered Saline (PBS) pH 7.4:
  - 80 g NaCl (1.37M)
  - 2 g KCl (0.03M)
  - 14.4 g Na_2HPO_4 (0.01M) or 18g $Na_2HPO_4 2H_2O$
  - 2.4 g KH_2PO_4
  - Made up to 1 litre with ddH_2O
- Sucrose Triton X-100 lysis Buffer:
  - 10 ml 1 M Tris-HCl (pH 8)
  - 5 ml 1M $MgCl_2$
  - 10 ml Triton X-100
  - Made up to 1 litre with ddH_2O
- T20E5:
  - 2ml 1 M Tris-HCl (pH 8)
  - 1 ml 0.5 M EDTA
- SDS (10%):

Add 10 g of SDS powder to 100 ml of ddH_2O

- Proteinase K solution:

  Add 25 ml of H_2O to stock powder (250 mg) to produce 10mg/ml

- Saturated NaCl:

  Add 40 g NaCl slowly to 100 ml of sterile water until saturated.

- TE:

  10 mM Tris-HCl pH 8.0 (1ml 1M)

  1 mM EDTA pH8.0 (0.2ml 0.5M)

  Made up to 100 ml with ddH_20.

2.2 Samples

This project is in collaboration with the Oesophageal Cancer research group led by Professor Iqbal Parker at the University of Cape Town, South Africa. This group includes a research nurse who recruits and consents participants and obtains blood samples. DNA was extracted from blood in the International Centre for Genetic Engineering and Biotechnology (ICGEB) at the University of Cape Town and an aliquot sent to King's College London (KCL).

Ethical approval for this study was obtained from the joint University of Cape Town/Groote Schuur Hospital Research Ethics Committee and the University of Stellenbosch/Tygerberg Hospital Ethics Committee.

For this study, patients with histologically confirmed primary invasive OSCC were recruited between March 2000 and August 2011 at Groote Schuur and Tygerberg Hospitals in Cape Town, South Africa. Controls were recruited from factories and outpatient clinics in the Western Cape between June 2000 and November 2010. Cases and controls were mainly from the South African Black and Mixed Ancestry populations, with other ethnicities excluded from the study. Over 98% of the Black cases and controls were Xhosa speakers.

Germ-line DNA from blood samples was initially available from 358 OSCC patients and 477 controls from the South African Black population, and 201 OSCC patients and 427 controls from the Mixed Ancestry population. During the course of the study, the sample size was expanded to a total of 407 OSCC patients and 849 controls from the Black population, and 257 OSCC patients and 860 controls from the Mixed Ancestry population.

DNA samples from matched blood-tumour pairs were available from 11 OSCC patients.

2.3 DNA extraction

2.3.1 Blood samples

DNA was extracted by collaborators at the University of Cape Town using the following method. Blood was transferred to sterile polypropylene tubes and diluted with 2 volumes of 1X PBS. Tubes were inverted to mix and centrifuged at 2200 g for 15 min. The supernatant was discarded and the pellet resuspended in 25 ml of Sucrose Triton X-100 lysis buffer and vortexed. Tubes were placed on ice for 5 min before centrifuging for 5 min at 2200 g. The supernatant was again discarded and the pellet re-suspended in 3 ml of T20E5 (0.6X the volume of original blood sample). To this, 300 µl 10% SDS was added together with 100 µl 10 mg/ml proteinase K. Tubes were inverted to mix and incubated overnight at 45ºC.

To each sample, 4 ml of saturated NaCl was added and mixed vigorously for 15 seconds. Tubes were then centrifuged for 40 min at 2400 g. To precipitate the DNA, 1 volume of absolute alcohol (at room temperature) was added and after gently agitating the tube, the DNA pellet was removed to an eppendorf tube. The pellet was washed in 1 ml of 70% ice-cold alcohol before centrifuging at 10000 rpm for 5 minutes. The alcohol was removed and the DNA pellet dissolved in 400 µl TE buffer.

2.3.2 Tissue samples

Tumour biopsy DNA was extracted by collaborators at the University of Cape Town using the QIAamp DNA Mini Kit. Briefly, the sample was brought to room temperature and then half of the tumour was homogenized in 80 µl of PBS in a 1.5ml microcentrifuge tube (the other half was used for RNA extraction, not used in this project). To each sample, 100 µl Buffer ATL was added followed by 20 µl Proteinase K. Samples were vortexed and incubated at 56°C, mixing occasionally, until the tissue was completely lysed. Then, 200 µl Buffer AL was

added and tubes pulse-vortexed for 15 sec before being incubated at 70°C for 10 min. Following this, 200 µl ethanol (96-100%) was added to the sample and again mixed by pulse-vortexing for 15 sec. The mixture (including any precipitate) was then applied to a QIAmp spin column (in a 2ml collection tube) and the tube (with closed cap) centrifuged at 6000 x *g* (8000 rpm) for 1 min. The QIAmp spin column was placed in a clean 2 ml collection tube and the tube containing the filtrate was discarded. Then, 500 µl Buffer AW1 was added to the QIAmp spin column and centrifuged at 6000 x g (8000 rpm) for 1 min. Again, the QIAmp spin column was placed in a clean 2 ml collection tube and the tube containing the filtrate discarded. To the QIAmp spin column, 500 µl Buffer AW2 was added and the tube centrifuged at full speed (20000 x *g*; 14000 rpm) for 3 min. The QIAmp spin column was placed in a clean 1.5 ml microcentrifuge tube and the collection tube discarded. Then, 200µl Buffer AE or distilled water was added to the tube and incubated at room temperature for 1 min. The tube was then centrifuged at 6000 x g (8000 rpm) for 1 min. The resulting filtrate contained the DNA. In order to obtain a higher yield, the QIAmp spin column was placed in a clean 1.5 ml microcentrifuge tube, and again, 200µl Buffer AE or distilled water was added to the tube and incubated at room temperature for 1 min. The tube was centrifuged at 6000 x g (8000 rpm) for 1 min and the DNA filtrate was combined with the rest of the DNA.

2.3.3 DNA quantification

DNA from the initial batches was quantified using Quant-iT™ PicoGreen ® dsDNA Reagent Kit according to manufacturer's instructions. The Picogreen reagent is a fluorophore and binds to double stranded DNA. Upon excitation, it is emits lights at a specified wavelength which is detected using a Tecan Genios Plate Reader.

DNA subsequently collected was quantified using Qubit® dsDNA BR (Broad-Range) Assay according to manufacturer's instructions, using 1:200 dilution of

DNA in the Qubit working solution. If the concentration was too low to be detected by this method, Qubit® dsDNA HS (High-Sensitivity) Assay was used, according to manufacturer's instructions. These Qubit reagents bind double stranded DNA and a Qubit Fluorometer detects the level of emission after excitation of the Qubit fluorophore.

2.4 Candidate gene case-control association studies

2.4.1 Polymerase chain reaction

2.4.1.1 Primer design

Primers were designed using Primer3 software (http://frodo.wi.mit.edu/primer3/) and UCSC In-Silico PCR was used to ensure primer pairs were specific to the region of interest (http://genome.ucsc.edu/cgi-bin/hgPcr). Primers were synthesized by Integrated DNA Technologies (USA).

2.4.1.2 Primer optimization

Polymerase chain reaction (PCR) conditions were optimized for each primer pair following the same standard method described here. Initially a temperature gradient was carried out which spanned a 10°C range around the annealing temperatures of the primers (typically 55°C – 65°C). The 10 µl PCR reactions contained 10 ng DNA, 1X PCR mastermix, 0.4 µM forward primer and 0.4 µM reverse primer. PCR was performed on a PTC-0225 DNA Engine (MJ Research), as were all other PCRs throughout this project, using the following conditions:

2 min at 92°C
20 sec at 92°C
30 sec at 55°C – 65°C } 30 cycles
30 sec* at 72°C
5 min at 72°C
Hold at 15°C

* Extension time depended on amplimer length, with every 500 base-pairs requiring an additional 30 seconds.

If the PCR product was not successfully amplified (via visualisation under UV light, see section 2.4.1.3), an alternative mastermix was made containing individual components for the reaction enabling quantities to be adjusted. For example, a typical 10 µl mastermix contained 10 ng DNA, 1X reaction buffer, 1.5mM $MgCl_2$, 0.4 µM forward primer, 0.4 µM reverse primer, 10% DMSO and 0.625 U GoTaq® Flexi DNA polymerase. If required, the magnesium chloride concentration was varied (1mM – 2mM $MgCl_2$) and DMSO removed.

2.4.1.3 Gel electrophoresis

Gel electrophoresis using DNA stained with ethidium bromide (EtBr) allows for visualization and sizing of a PCR product. Briefly, 2% agarose gels were made by dissolving 1g agarose in 50ml 1X TBE and, once cool to touch, 2.5 µl EtBr (10 mg/ml) was added and swirled to mix. The solution was poured into gel tanks with end-plates and combs, and left to cool and solidify for 20 minutes. Following removal of end-plates and combs, a running buffer of 50 ml 1X TBE containing 2.5 µl EtBr was poured into tank. To each DNA sample, 5 µl Orange G loading dye was added before loading into the gel wells. In an adjacent well, 5 µl 100bp DNA ladder was added to enable the size of the PCR product to be determined. A current of 100V was then applied to the gel tanks for an appropriate time to allow for resolution of DNA bands. Samples were visualized under UV light using a bench top UV transilluminator (UVP).

2.4.2 Genotyping assays

2.4.2.1 *CASP8* insertion/deletion genotyping

Genotyping of a 6-bp (AGTAAG) insertion/deletion in *CASP8* gene (-652 6N del, rs3834129) was achieved by determining the size of the PCR products. Primers were previously designed (Morgan *et al.* 2005; Sun *et al.* 2007), see Table 2.1. The forward primer was fluorescently labelled at the 5' end with 6-FAM.

Table 2.1: *CASP8* -652 6N del (rs3834129) primers

Variant	Forward primer	Reverse primer	Product length (bp)
CASP8 -652 6N del	CTGCATGCCAGGAGCTAAGT	GCCATAGTAATTCTTGCTCTGC	171 & 177

Briefly, the 5 µl PCR reaction contained 1X PCR mastermix, 0.4 µM of each primer and 20 ng of DNA. PCR conditions were as previously described (see p.70) using an annealing temperature of 62°C and an extension time of 30 seconds.

The PCR products were diluted 1:20 using Milli-Q H_2O and then 1 µl of this was added to a 96-well non-skirted plate. To each well, 10 µl of a LIZ loading cocktail (containing 9.88 µl HiDi Formamide and 0.12 µl GeneScan 500 LIZ size standard) was added. The samples were then heat denatured for 2 minutes at 95°C on a thermocycler. PCR products were separated by capillary electrophoresis on an ABI3730xl DNA Analyzer (Applied Biosystems) and sized using GeneMapper software (Applied Biosystems).

2.4.2.2 *PLCE1* insertion/deletion genotyping

The 14bp indel (CCCGGGCTCTGCCT) in the 5' untranslated region of *PLCE1* exon 1 was amplified and genotyped by size separation of the PCR products

using 3% agarose gel electrophoresis and visualised with ethidium bromide/UV light. Primers used for amplification are shown in Table 2.2.

Table 2.2: *PLCE1* 14 bp indel primers

Variant	Forward primer	Reverse primer	Product length (bp)
PLCE1 14bp indel	GGGAGCGGACTGTGAACG	GTGTCCCCGCTACTGTGTGT	217 & 203

The 10 µl PCR reaction contained 1X reaction buffer, 1.5 mM $MgCl_2$, 0.4 µM of each forward and reverse primer, 1 U GoTaq® Flexi DNA polymerase, 0.2 mM dNTP and 10% DMSO. PCR conditions were as previously described (see p.70) using an annealing temperature of 63°C and an extension time of 30 seconds.

2.4.2.3 TaqMan SNP genotyping assays

TaqMan SNP genotyping assays were designed and synthesized by Applied Biosystems. Table 2.3 shows the assay used for each variant genotyped. Custom assays were designed by Applied Biosystems after submitting the nucleotide sequence for the region of interest together with the alternative allele for the polymorphism. Primer and probe sequences for the custom assays are shown in the Appendix, Table A.3, but are not available for validated assays.

Table 2.3: TaqMan SNP genotyping assays

Variant	TaqMan genotyping assay		
	Drug metabolism assay	Validated SNP assay	Custom SNP assay
ADH1B rs1229984	✓		
ALDH2 rs886205	✓		
ALDH2 rs671	✓		
ADH7 rs1573496	✓		
COX-2 rs689466		✓	
FAS rs1800682		✓	
FAS rs2234767		✓	
FASL rs763110		✓	
MGMT rs12917		✓	
PLCE1 rs2274223		✓	
C20orf54 rs13042395		✓	
RUNX1 rs2014300		✓	
PDE4D rs10052657		✓	
Near *UNC5CL* rs10484761		✓	
COX-2 rs20417			✓
CASP8 rs1045485			✓
ALDH2 rs441			✓

Reactions were carried out according to manufacturer's instructions but used a 2.5 µl reaction volume with half the recommended amount of the TaqMan assay mix and with ABgene's Absolute QPCR ROX mix. Initially, 1µl DNA (20 ng/µl) was added to wells in a 96-well semi-skirted low-profile PCR plate (Starlab) and dried for 9 minutes at 65°C. A mastermix containing 1.25 µl ROX mix, 0.03125 µl TaqMan assay mix and 1.21875 µl Milli Q H_2O was made per reaction and added to wells. PCR was performed using one of the following sets of conditions.

For validated and custom TaqMan SNP genotyping assays:

15 min at 95°C
15 sec at 92°C
60 sec at 60°C } 45 cycles
Hold at 15°C

For TaqMan Drug Metabolism Genotyping Assays:

15 min at 95°C
15 sec at 92°C } 50 cycles
90 sec at 60°C }
Hold at 15°C

Fluorescent levels at the PCR end-point were determined using a 7900HT Fast Real-Time PCR system (Applied Biosystems) and genotypes assigned using SDS 2.2.2 software (Applied Biosystems).

2.4.2.4 KASPar SNP genotyping assays

The nucleotide sequence for the SNPs of interest and surrounding regions were submitted to KBioscience who designed and synthesized KASPar-by-design assays. The primer sequences are not made available. Reactions were carried out following the 'KASP version 4.0 SNP Genotyping Manual' in 4 µl reaction volumes. Initially, 1 µl 20 ng/µl DNA was dried to 96-well semi-skirted low-profile PCR plate and then a mastermix containing 0.055 µl SNP assay mix, 2 µl KASPar v4.0 High Rox reagent and 2 µl Milli Q H_2O was made per reaction and added to wells. The following conditions were used for PCR:

15 min at 94°C
20 sec at 94°C
60 sec at 65°C -57°C (dropping 0.8°C per cycle) } 10 cycles
20 sec at 94°C } 30 cycles
60 sec at 57°C }

Fluorescent levels at the PCR end-point were determined using a 7900HT Fast Real-Time PCR system (Applied Biosystems) and genotypes assigned using SDS 2.2.2 software (Applied Biosystems).

2.4.3 Sequencing of *PLCE1* exons

2.4.3.1 Amplification of *PLCE1* exons

PCR was carried out in a 10 µl reaction containing 10 ng DNA, 1X PCR mastermix, 0.4 µM of each forward and reverse primer for all reactions apart from amplification of exon 1. This 10 µl reaction contained 10 ng DNA, 1X reaction buffer, 1.5 mM $MgCl_2$, 0.4 µM of each primer, 0.2 mM dNTPs, and 10% DMSO. Primers for the amplification of the 34 exons of *PLCE1* are shown in Appendix, Table A.1, together with the annealing temperature and extension time for each primer set. PCR conditions were as previously described (see p.70).

2.4.3.2 Sanger sequencing of *PLCE1* exons

The PCR product was cleaned using ExoSAP-IT, where 1 µl PCR product was added to 0.25 µl ExoSAP-IT and 5.75 µl Milli Q H_2O. This was heated to 37ºC for 30 minutes followed by 15 minutes at 80ºC to inactivate the enzymes. Forward and/or reverse cycle sequencing was then performed in a 5.25 µl reaction volume using 3.5 µl of cleaned PCR product together with 0.25 µl of 20 µM sequencing primer (see Appendix, Table A.2) 0.25 µl of BigDye Terminator v3.1 and 1.25 µl of sequencing buffer. Reactions were run on a thermocycler with the following conditions:

30 sec at 96ºC
15 sec at 50ºC
1 min at 60ºC
} 30 Cycles
Hold at 10ºC

Purification of the sequencing product was then carried out to remove unincorporated dye terminators by ethanol precipitation. To the sequencing product, 20 µl precipitation solution was added and the plate sealed and briefly vortexed before incubating at room temperature for 15 mins. The plate was then centrifuged at 3000 rpm in a Sorvall Legend RT centrifuge for 30 mins.

Following this, plates were inverted onto blotting paper and briefly centrifuged (400 rpm for 10 seconds). To each well, 100 µl 70% ethanol was then added and the plate sealed and briefly vortexed. The plate was centrifuged at 3000 rpm for 10 mins, and again plates were inverted onto blotting paper and briefly centrifuged. The plate was dried at 60ºC for 5 minutes. To each well, 10 µl Hi-Di formamide was added and the plate sealed with a grey septa seal before being heat denatured at 95 ºC for 2 minutes. Sequencing products were then subjected to electrophoresis on a ABI 3730xl DNA sequencer (Applied Biosystems) and the results aligned to the human reference genome and analyzed using Staden software package.

2.4.4 Statistical analysis

2.4.4.1 Hardy-Weinberg equilibrium (HWE)

Hardy-Weinberg equilibrium (HWE) describes the principle that genotype frequencies are determined by the allele frequencies in a random mating population. Deviations from the expected frequencies can indicate problems with the data, such as genotyping errors, or inbreeding and population stratification. In disease groups, this can also signify association with the trait. If deviations are present in the control group, variants which fail to meet a specific threshold should be removed from analysis.

If alleles *A* and *a* have frequencies *p* and *q*, then under HWE genotypes *AA*, *Aa* and *aa* will have frequencies p^2, $2pq$ and q^2, respectively. The genotype frequencies must also sum to 1, hence:

$$p^2 + 2pq + q^2 = 1$$

The expected genotype frequency can be calculated based on the experimental allele frequencies. The frequencies of the observed (O) and expected (E) genotypes can then be tested for a significant difference using a chi squared (χ^2) test with one degree of freedom:

$$\chi^2 = \sum_i (O_i - E_i)^2 / E_i$$

HWE chi-squared values were calculated using Microsoft Excel, with these values converted to a p-value using the 'CHIDIST' function. P <0.05 was considered as a statistically significant deviation from the HWE for candidate gene association tests.

2.4.4.2 Association tests

Genotype and allele frequencies were calculated for cases and controls. To determine if there was a significant difference between these two groups, the minor allele frequencies (both allelic and genotypic) were compared using a Pearson's chi-squared (χ^2) test, which was performed using a two-by-two table (http://www.quantitativeskills.com/sisa/statistics/twoby2.htm). For a single test, a threshold of P< 0.05 was used to identify a statistically significant difference.

When multiple testing occurred, the Bonferroni correction was applied. This defines the significance level as 0.05/*n*, where *n* is the number of tests performed. In Chapter 3, the significance threshold was set at P<0.0042 (0.05/12) for the testing of 12 variants that were polymorphic in the South African populations to ensure the overall type I error rate was no more than 5%. In Chapter 4, a significance threshold of P<0.01 (0.05/5) was used for the association tests of the five SNPs that were identified in the Chinese genome-wide association studies. This threshold was also applied to the association tests of the five additional *PLCE1* variants. No additional correction was applied for the two populations tested.

2.4.4.3 Odds ratio

The odds of an event occurring is a ratio of the probability that an event occurs to the probability that it does not. The odds ratio (OR) is used to estimate the

risk of a particular outcome, and is the ratio of the odds of an event occurring in those with an outcome of interest, e.g. cancer patient, against those without the outcome, e.g. controls. For example, if *A* and *B* represent the number of cancer cases with the alternative or a common reference alleles, respectively, and *C* and *D* represent the number of controls with the alternative or a common reference alleles, respectively, then:

$$OR = (A/B) / (C/D)$$

An odds ratio of greater than one indicates that the event is more likely to happen (e.g. allele is associated with increased risk of disease). If the odds ratio is less than one, the event is less likely to happen and can indicate a protective effect. If the odds ratio equals one, then events are just as likely to occur and show no association. For each OR, 95% confidence intervals (CI) are also reported and are calculated by the formula:

$$95\%\ CI = \exp(\log(OR) \pm 1.96\ SE(\log OR))$$

where

$$SE(\log OR) = \sqrt{\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}}$$

is the standard error (SE) of the log odds ratio.

Genotypic and allelic odds ratios together with 95% confidence intervals were calculated using the common homozygous genotype or common allele as the reference. These were determined using the formulas described above in Microsoft Excel.

2.4.4.4 Linkage disequilibrium

Linkage disequilibrium (LD) is the non-random association of alleles at two or more loci (Slatkin 2008). The level of LD depends on factors such as natural

selection, genetic drift, non-random mating, the amount of recombination and the rate of mutation.

LD is commonly measured using D' and r^2 values. Both of these derive from D, which is the the difference between the frequency of loci carrying the pair of alleles A and B at two loci (P_{AB}) and the product of the frequencies of those alleles (P_A and P_B):

$$D_{AB} = P_{AB} - P_A P_B$$

D' is the ratio of D to its maximum possible value, the smaller of $P_A(1-P_B)$ and $P_B(1-P_A)$, given the allele frequencies (Lewontin 1964). Alternatively, r^2 is the correlation coefficient of the allele frequencies and is calculated by:

$$r^2 = \frac{D^2}{P_A(1-P_A)P_B(1-P_B)}.$$

Both D' and r^2 take values between 0 and 1 which show linkage equilibrium and complete/perfect linkage disequilibrium, respectively. The former implies that the loci are independent of each other, whereas the latter is true only if alleles have not been separated by recombination. An r^2 of 1 also requires the SNPs to have the same allele frequencies, but this is not the case for D'.

Determination of LD between variants in the same gene was performed using UNPHASED v3.1.5 (Dudbridge 2008) (for *COX-2*, *FAS*, *ALDH2* and *CASP8*) and Haploview (for *PLCE1* variants) (Barrett *et al.* 2005).

2.4.4.5 Haplotype analysis

Haplotype analysis was performed using UNPHASED v3.1.5 (Dudbridge 2008) when multiple variants in the same gene had been genotyped. Haplotype frequencies were determined, followed by haplotype case-control association tests using the most common haplotype as the reference. A full likelihood model

was used for analysis, testing the null hypothesis that all the haplotype odds ratios are equal.

2.4.4.6 Gene-environment interactions

SNPs that showed a suggestive allelic association with OSCC ($P<0.05$) were further investigated for the effect of alcohol and tobacco by stratifying cases and controls by smoking and alcohol drinking status and performing three association tests. An example of these tests is shown here for alcohol use: a case-only analysis (drinkers vs. non-drinkers), a case-control analysis for drinkers only, and a case-control analysis for non-drinkers only. In addition, tests of gene-environment interaction were performed using logistic regression in PLINK, adjusting for age, gender, alcohol and tobacco consumption. An example for gene-alcohol interactions is shown below:

$$Y = \beta_0 + \beta_1 ADD + \beta_2 Alcohol + \beta_3 Tobacco + \beta_4 Age + \beta_5 Gender + \beta_6 ADDAlcohol$$

Where $Y=\log(p/1-p)$ is the logit of the probability (p) of being a case, β_0 is a constant, β_1- β_5 are the SNP effects adjusted for each of the covariates and β_6 is the interaction term to be tested with the null hypothesis H_0: $\beta_6=0$.

In Chapter 3, demographic data for controls only became available after the study was published (Bye *et al.* 2011), and therefore, the gene-environment analyses are included as a supplementary analysis here.

2.4.4.7 Power

The power of a study, that is the probability that a specific magnitude of odds ratios can be detected given allele frequency and sample size, was determined using Quanto (http://hydra.usc.edu/gxe/) (Gauderman 2002).

2.5 Case-control association study using the Immunochip

2.5.1 Samples

A subset of samples from the South African Black and Mixed Ancestry populations were chosen for genotyping on the Immunochip. Samples with the highest DNA yield were selected as well as ensuring a spread across the recruitment period to ensure all collection phases were represented. Sample numbers were determined by funding available. From the Black population, 300 cases and 300 controls were genotyped for a case-control association study and additionally, 50 cases and 50 controls from the Mixed Ancestry population were selected to provide insight into their population structure.

2.5.2 Genotyping

Genotyping was carried out at the Genomics core facility, part of the NIHR Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London.

Samples were genotyped using the Infinium® HD Assay Ultra Protocol (Catalog # WG-901-4005; Part # 11328095 Rev. B), following manufacturer's instructions. Briefly, 200 ng DNA per sample was denatured and neutralized before overnight amplification. DNA was then fragmented by an enzymatic reaction and precipitated using isopropanol. The DNA was resuspended and hybridized to the BeadChip, with non-hybridized DNA being washed away. The oligonucleotides on the array were then extended by a single nucleotide base using the sample DNA as a template. The newly incorporated bases contain a fluorescent label which can be excited using a laser. This process was carried out in the Illumina iScan machine which then detects and records the light emitted.

2.5.3 Genotype calling and quality control

GenomeStudio Data Analysis Software (Illumina) was used for visualizing genotype clusters and for the initial genotype calling and quality control (QC). At this stage, samples which had a genotyping call rate considerably lower than all other samples (<90%) were removed.

Data for the remaining samples were then re-formatted for use with optiCall (http://www.sanger.ac.uk/resources/software/opticall/). OptiCall is a genotype calling software, which uses both within and across sample intensity information to identify intensity regions that have a high probability to contain each of the three genotypes (Shah *et al.* 2012). Using this approach, rare variants can be called if their intensity is in a high probability region. This work was carried out by Dr. Sarah L. Spain.

2.5.3.1 Population stratification

The ethnicity of study participants is self-declared and recorded in a database. Principal components analysis (PCA) can be used to detect population stratification and, hence, identify individuals who are outliers from the self-declared population. For this analysis, only SNPs that were independent of each other were used, which was based on those with a pairwise LD of $r^2<0.2$ and a minor allele frequency of greater than 5%. This amounted to a total of 27,132 variants. Altogether, 661 samples were included in the analysis from both South African populations.

Dr. Sarah Spain performed PCA using EIGENSTRAT. Samples which clustered with a different population (e.g. those self-declared as from the Black population but who clustered with the Mixed Ancestry population in the PCA) were reassigned to the population determined by the PCA. Population outliers, samples which did not cluster with individuals from either the Black or Mixed Ancestry populations, were also present (see Chapter 5, section 5.3.1.1 for the

thresholds used based on the PCA plots). Sample relatedness between outliers was determined (see below), and one from each pair of highly related samples was removed from the analysis.

2.5.3.2 Further quality control

PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/) was used for QC with the majority of steps performed by Dr. Sarah Spain (indicated below). The following QC steps were performed:

- The percentage of SNPs that were successfully assigned a genotype for each individual (sample genotype call rates) was determined, and those samples with a call rate of <95% were removed. This was determined using the PLINK command "-- missing" (Dr Spain).
- Related samples were identified based on PI-Hat scores of >0.5 (using the PLINK command '-- genome'), and one from each pair (the one with the highest level of missing data) was removed (Dr Spain).
- The percentage of samples successfully assigned a genotype for each SNP (the SNP call rate) was calculated (using the PLINK command "-- missing") and SNPs with a call rate of <95% were removed (Dr Spain).

Additionally, cases and controls were mainly genotyped on separate plates and hence, to determine if plate effects were present, minor allele frequencies for SNPs with P<0.05 in the association test, were compared for each plate individually against all other plates using Q-Q plots. This allows plates with an unusual distribution of minor allele frequencies to be identified.

2.5.4 Case-control genetic association study

After the QC described above, 278 cases and 257 controls (see Table 2.4) with genotype data for 139,793 SNPs were available for the case-control analysis in the South African Black population. SNPs which had a significant deviation (P<

1 x 10^{-6}) from HWE in controls were excluded (based on the Bonferroni correction, see below). Allele frequencies in cases and controls were compared using the '- -assoc' function in PLINK, which is an allelic chi-square test with 1-degree-of-freedom, with odds ratios and P-values calculated. The Bonferroni correction was used to account for the multiple testing of SNPs. This was based on 27,132 independent SNPs present (pairwise LD of r^2<0.2 and MAF >0.05), which gave a significance threshold of P<1.84 x 10^{-6} (0.05/27,132).

Table 2.4: Demographic information for South African Black population samples used in Immunochip analysis

	Cases	Controls
Total	278	257
Age, mean years (SD)	60.3 (11.5)	48.1 (17.5)
Sex, *n* (%):		
Male	134 (48.2%)	91 (35.4%)
Female	144 (51.8%)	164 (63.8%)
Unknown	0	2 (0.8%)
Smoking status, *n* (%):		
Smoker	159 (57.2%)	96 (37.4%)
Non-smoker	119 (42.8%)	158 (61.5%)
Unknown	0	3 (1.2%)
Alcohol consumption, *n* (%):		
Drinker	176 (63.3%	128 (49.8%)
Non-drinker	100 (36.0%)	128 (49.8%)
Unknown	2 (0.7%)	1 (0.4%)

2.5.4.1 Association plots

For SNPs with an association with OSCC of P<1 x 10^{-5}, the p-value for this and variants in the surrounding region (~1000kb) were visualized using SNAP (SNP annotation and proxy search) regional association plots (http://www.broadinstitute.org/mpg/snap/ldplot.php). In addition, pairwise LD (r^2) for the index SNP compared to all other variants, as determined using PLINK, were also visualized on this plot.

2.5.4.2 Extension study

SNPs with an association with OSCC of $P_{Immunochip} < 1 \times 10^{-4}$ were considered for this larger case-control extension study, with variants selected based on the SNAP association plots (SNPs were prioritized if the association was supported by nearby variants with p-values of similar magnitude) and LD levels (pairs of SNPs in high LD were not both genotyped). Seven variants were chosen for TaqMan genotyping in the South African Black population (see Table 2.5), with the SNPs genotyped in all available cases and controls.

Table 2.5: Variants genotyped in Immunochip extension study

Variant	Chr	Position (b37)	Gene	Location
rs9887787	1	92222143	*TGFBR3*	Intronic
rs2810893	1	92144970	*TGFBR3*	Downstream
rs2182833	1	55500429	*PCSK9*	Upstream
rs13390918	2	199564895	*Intergenic*	-
rs13147507	4	115334709	*Intergenic*	-
rs7714035	5	102644627	*Intergenic*	-
rs36590	22	30328070	*MTMR3*	Intronic

A well-powered replication using an independent sample set was not possible due to inadequate numbers of OSCC cases, and hence the samples genotyped by Immunochip were also included in the TaqMan genotyping assay. In total, an additional 126 cases and 577 controls were available, giving a total of 404 OSCC cases and 834 controls for genotyping in this extension study (see Table 2.6). Samples identified in the PCA analysis as not being from the Black population (7 cases and 16 controls) were not genotyped, and the self-declared Mixed Ancestry samples which were shown to cluster with the Black population in the PCA plot were included (6 cases and 2 controls). The SNPs were genotyped using validated TaqMan SNP assays, as described in section 2.4.2.3.

Table 2.6: Demographic information for samples used in the Immunochip extension study

	Cases	Controls
Total, *n*	404	834
Age, mean years (SD)	59.8 (11.2)	48.7 (16.8)
Sex, *n* (%):		
Male	199	329
Female	205	502
Unknown	0	3
Smoking status, *n* (%):		
Smoker	240	327
Non-smoker	163	496
Unknown	1	11
Alcohol consumption, *n* (%):		
Drinker	252	444
Non-drinker	149	386
Unknown	3	4

2.5.4.3 Gene-environmental interactions

Alcohol and tobacco effects were investigated as described on page 81. Additionally, an Immunochip-wide gene-environment interaction analysis was performed, based on the genome-wide gene-environment analysis by Wu *et al.* (2012.a). In this method, the interaction between each variant on the Immunochip and drinking or smoking status was tested using logistic regression with age, gender and alcohol and smoking status as covariates using PLINK, as described on page 81.

2.6 Somatic mutation identification using whole-exome sequencing

Matching blood and tumour DNA were whole-exome sequenced in 8 OSCC patients to identify somatic mutations. This project was in collaboration with Dr Iwanka Kozarewa and colleagues at the Institute of Cancer Research (ICR), London, who had developed a protocol to sequence low quantities of input DNA (>50 ng), compared to the standard Agilent protocol of 3 µg DNA. Six tumour samples had a low DNA yield (500 ng) and, hence, were sequenced at the ICR using this adapted method. The matching blood DNA was also sequenced at the ICR using the standard protocol. Table 2.7 shows the protocol used and the location of sample preparation and sequencing.

Table 2.7: Whole-exome sequencing protocol list

Sample	Exome-sequencing protocol	Sample preparation		Sequencing		
		KCL	ICR	KCL	ICR	Illumina, USA
P662	A	✓				✓
232T	A	✓				✓
P1282	B	✓		✓		
386T	B	✓		✓		
P1377	B	✓		✓		
T437	B	✓		✓		
P1354	B		✓		✓	
T416	C		✓		✓	
P1408	B		✓		✓	
T443	C		✓		✓	
P1400	B		✓		✓	
T438	C		✓		✓	
P1406	B		✓		✓	
T442	C		✓		✓	
P1116	B		✓		✓	
T441	C		✓		✓	

Protocol A: SureSelect Target Enrichment System for Illumina Paired-End Sequencing Library SureSelect Human All Exon and Human All Exon Plus Protocol Version 2.0.1, May 2010; **Protocol B**: SureSelectXT Target Enrichment System for Illumina Paired-End Sequencing Library. SureSelectXT Target Enrichment for Illumina Multiplexed Sequencing Version 1.2, May 2011;
Protocol C: ICR adapted protocol for low-input DNA.
KCL: King's College London; **ICR:** Institute of Cancer Research

2.6.1 Sample preparation

2.6.1.1 Standard protocol

The standard sample preparation used Agilent's "SureSelectXT Target Enrichment System for Illumina Paired-End Sequencing Library (SureSelectXT Target Enrichment for Illumina Multiplexed Sequencing), Version 1.2, May 2011". Two samples used an older version of this protocol, "SureSelect Target Enrichment System for Illumina Paired-End Sequencing Library (SureSelect Human All Exon and Human All Exon Plus), Version 2.0.1, May 2010", which mainly differed in the ability to add index tags. Therefore, the latest method will briefly be described, which is also summarized in Figure 2.1.

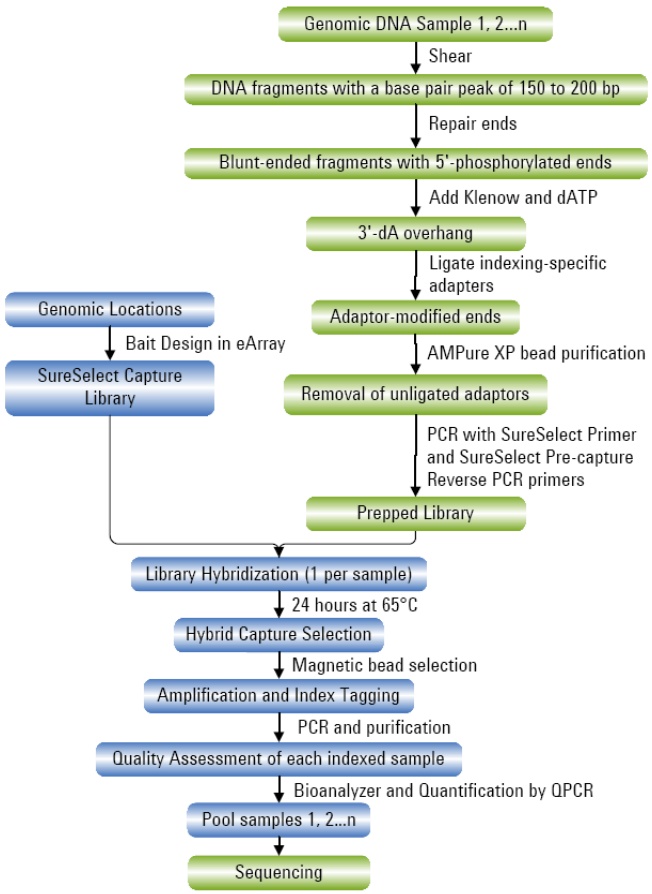


Figure 2.1: Agilent's sample preparation for whole-exome sequencing (taken from Agilent's protocol)

DNA (3 µg) was initially sheared to 150-200 nucleotides in length using the Covaris system. The ends of the DNA were repaired to produce blunt ended fragments, with 'A' bases added to the 3' end. Adaptors were then ligated to the DNA and the resulting DNA library amplified using PCR. Between each of these steps, samples were purified using Agencourt AMPure XP beads to remove non-specific DNA. Next, the prepped DNA library was hybridised to the whole-exome capture library. Hybridised DNA was selected using streptavidin coated magnetic beads, which bind the biotinylated capture library baits, with unbound DNA discarded. The hybridised library was then amplified with index tags added. Samples were then prepared for cluster amplification and run on the Illumina HiSeq 2000. For samples sequenced at King's College London, this final step was performed by staff at the Genomics Core facility.

2.6.1.2 ICR low-input DNA protocol

For samples with a low yield of DNA (<3 µg), the standard protocol (Agilent's "SureSelectXT Target Enrichment System for Illumina Paired-End Sequencing Library (SureSelectXT Target Enrichment for Illumina Multiplexed Sequencing), Version 1.2, May 2011") was adapted by Iwanka Kozarewa *et al.* at the ICR (personal communication). For each tumour sample, 500 ng starting DNA was used, although further adaptations of the protocol could be performed in order to sequence DNA from as little as 50ng. The method used to sequence 500 ng starting DNA is briefly compared to the standard protocol in Figure 2.2. The full low-input DNA protocol is described following this, although where long steps are identical to the standard protocol, readers are directed to this.

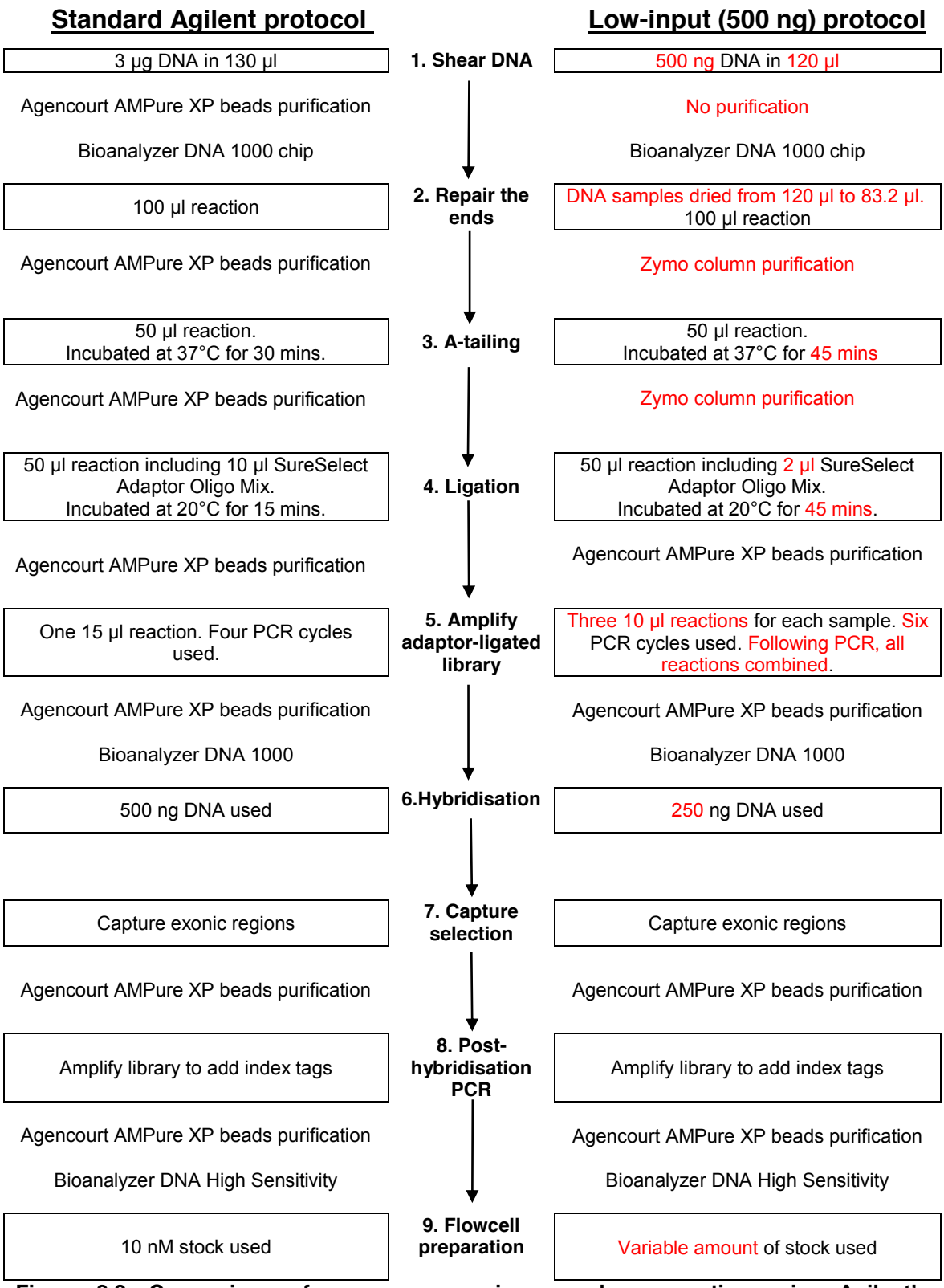


Figure 2.2: Comparison of exome sequencing sample preparation using Agilent's protocol vs. low-input protocol

The low input protocol used 500 ng starting DNA. Differences in the protocols are highlighted in red. (Adapted from Kozarewa *et al.* 2012)

1. Shear DNA
   - 500 ng DNA was diluted with 1x TE buffer to a total volume of 120 µl in a 1.5 ml microcentrifuge tube.
   - DNA was transferred to a Covaris microTube and samples sheared using a Covaris instrument using the following conditions:

Duty Cycle:	10%
Intensity:	5
Cycles per Burst:	200
Time:	6 cycles of 60 seconds each
Set Mode:	Frequency sweeping
Temperature:	4 – 7°C

   - Sheared DNA was transferred into a clean 1.5 ml microcentrifuge tube.
   - DNA quality was assessed with a Bioanalyzer DNA 1000 chip with fragments expected to be 150-200 nucleotides.

2. Repair the ends
   - DNA was dried down to 83.2 µl or below. If less, nuclease-free water was added to make up the volume.
   - For each sample, the end repair reaction contained the following reagents and were mixed well by pipetting:

DNA + nuclease-free water:	83.2 µl
10X End repair Buffer:	10 µl
dNTP mix:	1.6 µl
T4 DNA polymerase:	1 µl
Klenow DNA polymerase:	2 µl
T4 polynucleotide kinase:	2.2 µl
Total:	100 µl

- Samples were incubated in a thermal cycler for 30 minutes at 20°C (with no heated lid).
- End-repaired DNA was purified using a Zymo DNA Clean & Concentrator™-5 column following manufacturer's instructions, except that 5 volumes of DNA Binding Buffer was added to each volume of DNA (500 µl DNA Binding Buffer to 100 µl DNA). After the washing step, DNA was eluted in 20 µl EB buffer (Qiagen) pre-warmed to 50°C.

3. A-tailing
   - After elution, ~15 µl DNA remained.
   - The A-tailing reaction contained the following reagents for each sample and were mixed well by pipetting:

DNA + nuclease-free water:	41 µl
10X Klenow polymerase buffer:	5 µl
dATP:	1 µl
Exo(-) Klenow:	3 µl
Total:	50 µl

   - Samples were incubated in a thermal cycler for 45 minutes at 37°C (with heated lid not exceeding 50°C).
   - A-tailed DNA was purified using a Zymo DNA Clean & Concentrator™-5 column following manufacturer's instructions, except that 5 volumes of DNA Binding Buffer was added to each volume of DNA (250 µl DNA Binding Buffer to 50 µl DNA). After the washing step, DNA was eluted in 15 µl EB buffer (Qiagen) pre-warmed to 50°C.

4. Ligation
   - After elution, ~11.5 µl DNA remained.
   - For each sample, the ligation reaction contained the following reagents and were mixed well by pipetting:

DNA + nuclease-free water:	36.5 µl
5X T4 DNA ligase Buffer:	10 µl
SureSelect Adaptor Oligo Mix:	2 µl
T4 DNA Ligase:	1.5 µl
Total:	50 µl

- Samples were incubated in a thermal cycler for 45 minutes at 20°C (no heated lid).
- Samples were purified using Agencourt AMPure XP beads, according to the standard protocol.
- A sub set of samples were run on Bioanalyzer High Sensitivity chip to confirm that DNA was present.

5. Pre-hybridisation PCR
   - After purification, ~32 µl DNA remained. From this, three pre-hybridisation PCR reactions were performed for each sample, each containing 10 µl DNA. Each reaction contained the following reagents and were mixed well by pipetting:

Indexing adaptor-ligated library (DNA):	10 µl
Nuclease-free water:	26 µl
SureSelect primer:	1.25 µl
SureSelect ILM indexing pre capture PCR reverse primer:	1.25 µl
5X Herculase II Rxn buffer:	10 µl
100 mM dNTP mix:	0.5 µl
Herculase II fusion DNA polymerase:	1 µl
Total:	50 µl

- Reactions were performed using the following thermocycler conditions:

  2 mins at 98°C
  30 sec at 98°C
  30 sec at 65°C } 6 cycles
  1 min at 72°C
  10 min at 72°C
  Hold at 4°C

- Following PCR, the three reactions for each sample were pooled (a total of 150 µl).
- Samples were purified using Agencourt AMPure XP beads, according to the standard protocol, except that 270 µl beads were added to DNA due to the increased volume of DNA (150 µl). An increased volume of ethanol was also used during the wash steps, if needed, to ensure beads were covered.
- DNA quality was assessed using a Bioanalyzer DNA 1000 chip and the concentration measured. Peak size should be approximately 250-275 bp.

6. Hybridization
   - After purification, ~30 µl DNA remained. The hybridization reaction required 250 ng DNA in a 3.4 µl volume so the samples were concentrated to achieve this.
   - The hybridization reaction was performed according to the standard protocol.

7. Capture selection
   - This step was performed according to the standard protocol, followed by purification using Agencourt AMPure XP beads as per instructions.

8. Addition of index tags by post-hybridization amplification
    - The reaction was performed according to Agilent's instructions, followed by purification using Agencourt AMPure XP beads as per instructions.
    - DNA quality was assessed using a Bioanalyzer High Sensitivity DNA chip and the concentration determined using qPCR, according to the standard protocol.

9. Pool samples for multiplexed sequencing
    - Four samples were pooled into one lane for sequencing.
    - The standard Agilent protocol used a final concentration of 10 nM for all DNA in the pool. For this adapted protocol, the final concentration was based on the lowest concentration of the samples in the pool. For example, if one sample had a concentration of 5 nM, then this concentration was used.
    - Four DNA libraries were combined so each sample was present at an equal amount to make a final concentration of 10 nM (or lower, see above). Elution buffer was added so the final volume was 19 µl.

10. Preparation for cluster amplification
    - To the 19 µl of DNA library, 1 µl 0.5N NaOH was added and then libraries were denatured by incubating for 5 minutes at room temperature.
    - The remaining preparation steps and the cluster generation followed the standard protocol.

2.6.2 Analysis pipeline

For samples prepared at KCL, sequences were aligned to the reference genome using Novoalign (novocraft.com) and PCR duplicates removed. Variants were called using SamTools (Li *et al.* 2009.b), and annotated using Annovar (Wang *et al.* 2010.b). These steps were performed by Dr Michael Simpson, KCL.

For samples sequenced at ICR, BWA (http://bio-bwa.sourceforge.net/bwa.shtml) was used to align sequences to the reference genome and PCR duplicates removed. Variants were called using GATK Broad Best pipeline (http://www.broadinstitute.org/gatk/index.php) with standard settings, and annotated using the Ensembl variation database (v61) (http://www.ensembl.org/index.html). Common variants (MAF of >5% in dbSNP) were removed. These steps were performed by Dr James Campbell, ICR.

The data for all variants that differed between blood and tumour DNA was then exported into a Microsoft Excel format where further thresholds could be set to identify mutations that were likely to be somatic.

2.6.3 Somatic mutation calling

A somatic mutation was called when the alternative allele was supported by ≥15% of reads in the tumour DNA but was absent in blood DNA (<2% to allow for sequencing errors). This threshold allows for ~33% of the tissue to be derived from tumour DNA, with the rest being normal tissue DNA. For example, if a heterozygous mutation arose in tumour DNA, the alternative allele would be expected to be present ~50% if the sample was pure tumour. If the tissue was derived from equal amounts of tumour and normal DNA, heterozygous somatic mutations would now appear to be 75% reference allele and 25% alternative allele.

If a very low number of somatic mutations were called using this threshold, it might indicate a high level of normal tissue contamination (>67%) in the tumour sample, and not simply a lack of mutations. Therefore, in these instances, the threshold was lowered to ≥10% of reads in tumour DNA. This was applied to two blood-tumour pairs (232T-P662 and T416-P1354).

In addition, the total number of reads (both reference and alternative alleles) required was ≥8 for blood DNA and ≥14 for tumour DNA. These thresholds were selected based on other published exome sequencing studies (Stransky *et al.* 2011; Dulak *et al.* 2013).

2.6.4 Sanger sequencing to confirm somatic mutations

Somatic mutations which were potentially functional (stop/gain, frameshifts and essential splice-site variants) or were non-synonymous mutations present in genes which were known to be mutated in cancer (according to the Cancer Gene Census - http://cancer.sanger.ac.uk/cancergenome/projects/census/) were prioritized for further work. Also prioritised were non-synonymous mutations in genes that were recurrently mutated. All of these mutations were visualised on Integrative Genomics Viewer (IGV) to confirm that they appeared to be valid (mutations present in the tumour with high confidence, but absent in the blood). Mutations which did not look convincing either contained variants in the normal DNA, or the region contained many SNPs or insertions/deletions called with varying degree of confidence, usually both in the tumour and normal DNA. Those that were valid mutations were then Sanger sequenced in the blood and tumour DNA to confirm the presence of the mutations using the standard method described on p.76, apart from a mutation in *JUNB*, which used the alternative mastermix with 10% DMSO. The primers and PCR conditions are shown in the Appendix, Table A.4 and Table A.5.

2.6.5 *TP53* exon amplification and sequencing

The exons of *TP53* were sequenced in 10 blood-tumour pairs, the total number of samples available (see Table 2.8). The *TP53* exons selected for Sanger sequencing were based on the IARC *TP53* database (http://p53.iarc.fr/Default.aspx - database version R16, November 2012) (Petitjean *et al.* 2007). As exon 1 and the majority of exon 11 are non-coding, the IARC sequencing protocol does not sequence these regions

(http://p53.iarc.fr/ProtocolsAndTools.aspx), and hence, were not sequenced in the blood-tumour pairs. Exons 2-11 were sequenced using the standard PCR protocol, see p.70. Primers and PCR conditions are shown in the Appendix, Table A.6. Exons were then sequenced according to the method previously described (see p.76), using the primers used for PCR.

Table 2.8: Samples used for *TP53* Sanger sequencing

Tumour	Blood
T416	P1354
T443	P1408
T438	P1400
T442	P1406
T441	P1116
GSHT	P1508
TBHT	TB62
T437	P1377
T386	P1282
232T	P662

2.6.6 *PPM1D* exon amplification and sequencing

The seven exons of *PPM1D* (transcript ENST00000305921) were sequenced in 11 blood-tumour pairs, the total number of samples available (see Table 2.8, plus 288T-P920 pair). PCR was carried out in 10 µl reactions. Exons 1, 3 and 4 contained 10 ng DNA, 1x reaction buffer, 1.5m $MgCl_2$, 0.2 mM dNTP, 0.4 µM forward primer, 0.4 µM reverse primer, 10% DMSO and 0.625 U Taq polymerase. Exon 2 mastermix contained 10 ng DNA, 1x reaction buffer, 1.5 mM $MgCl_2$, 0.2 mM dNTP, 0.4 µM forward primer, 0.4 µM reverse primer and 0.625 U Taq polymerase. Exon 5+6, and the four amplicons of exon 7 used the standard PCR mastermix as described on p.70. Primers and PCR conditions are shown in the Appendix, Table A.7. *PPM1D* exons were then sequenced according to the method previously described (see p.76), using the primers used for PCR.

3 Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa

The work described in this chapter was published in 2011 in *Carcinogenesis* (Bye *et al.* 2011). The published paper is inserted into the thesis on pages 112-121. A more detailed background to the work will first be given.

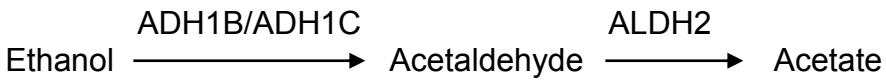
The main aim of this chapter was to investigate genes and variants that had previously been reported to be associated with OSCC, and to determine whether they were also risk factors in South African populations. Genes and variants with robust evidence of association in multiple studies were prioritised for these case-control genetic association studies.

3.1 Candidate genes and OSCC

Before the advent of genome-wide association studies, one way of identifying disease susceptibility loci was through candidate gene association studies. In oesophageal cancer studies, genes involved in apoptosis, cell proliferation and DNA repair are good candidates as deregulation of these processes can cause uncontrolled growth leading to cancer development. Additionally, alcohol and tobacco smoking are known risk factors for OSCC and, therefore, genes involved in alcohol metabolism and tobacco detoxification are potentially important in OSCC development. These genes, together with evidence of their association with OSCC susceptibility, will now be discussed.

3.1.1 Alcohol metabolism

Upon consumption, alcohol is metabolized to acetaldehyde within the liver by alcohol dehydrogenase (ADH) enzymes, primarily ADH1B and ADH1C, where it is then oxidised to harmless acetate mainly by aldehyde dehydrogenase 2 (ALDH2). This is summarized in below:



3.1.1.1 Aldehyde dehyrogenases

Variants in *aldehyde dehydrogenase 2* (*ALDH2*) are well known genetic risk factors for OSCC. In particular, the *ALDH2* Glu504Lys (rs671) SNP has been associated with an increased susceptibility to OSCC in a number of studies in Japanese and Chinese populations (Yokoyama *et al.* 2002; Cui *et al.* 2009; Tanaka *et al.* 2010). The functional effect of *ALDH2* Glu504Lys has been characterised, with lysine at this position shown to result in a catalytically inactive ALDH2 subunit which is unable to oxidise acetaldehyde (Yoshida *et al.* 1984). Hence, lys/lys homozygotes have no enzyme activity and heterozygotes only have ~6% of the normal ALDH2 activity (Crabb *et al.* 1989; Yoshida *et al.* 1991). This large enzymatic reduction observed in heterozygotes is due to ALDH2 being a tetrameric protein and the presence of lysine even in a heterozygous form affects the stability of the tetramer structure (Oota *et al.* 2004). The ALDH2 lysine residue, therefore, acts in a dominant manner (Crabb *et al.* 1989).

The Glu504Lys polymorphism is almost unique to Asian populations (Li *et al.* 2009.a), with allele frequencies varying from 16-26% for the HapMap Chinese and Japanese populations (http://hapmap.ncbi.nlm.nih.gov/index.html.en). It is this variant that is responsible for the alcohol flushing response (“Asian flush”), whereby facial reddening and nausea occurs after alcohol consumption. This reaction, therefore, should potentially protect against alcohol-related diseases as Lys/Lys carriers will avoid alcohol. Indeed, this has been shown to be the case in oesophageal cancer studies; the odds ratio for developing this cancer is 0.36 (95% CI = 0.16 – 0.80) for *ALDH2* Lys/Lys individuals compared to Glu/Glu individuals who have normal enzymes activity (Lewis and Smith 2005; Cui *et al.* 2009). Heterozygous individuals are the group most at risk of developing OSCC as they have a less severe flushing reaction than Lys/Lys homozygotes, to

which they may develop tolerance, but may consume large amounts of alcohol (Brooks *et al.* 2009.a).

The *ALDH2* 504Lys allele is thought to have originated in the Han Chinese population of Central China but the reason for the high allele frequency there is unknown (Li *et al.* 2009.a). It is possible that ALDH2 is involved in other cellular pathways in which the 504Lys allele is beneficial, thus outweighing the disadvantage of acetaldehyde accumulation after alcohol consumption (Oota *et al.* 2004). Alternatively, Oota *et al.* propose that higher acetaldehyde concentrations may inhibit growth of parasites and hence, provide resistance to diseases.

In addition to acetaldehyde levels being increased in the liver, it has been observed that the levels of this compound are also greatly increased in saliva due to the ability of oral bacteria to oxidise ethanol to form acetaldehyde but with limited capability to form acetate (Toh *et al.* 2010). Acetaldehyde levels can be 10-100 times greater in the saliva than in blood and the direct contact of acetaldehyde in the saliva with the oesophagus may be another risk factor for the development of OSCC (Toh *et al.* 2010).

Other *ALDH2* variants have also been associated with an altered susceptibility to OSCC, including in European populations in whom the *ALDH2* 504Lys allele is absent. These polymorphisms include *ALDH2* +82A>G, +348 C>T and -261 C>T (Hashibe *et al.* 2006). The functional significance of these polymorphisms is not clear, although if they affect the stability of the ALDH2 protein, it may lead to abnormally high levels of acetaldehyde after drinking alcohol.

3.1.1.2 Alcohol dehydrogenases

Several alcohol dehydrogenase (ADH) genes have been associated with OSCC and other upper aerodigestive cancers, including *ADH1B*, *ADH1C* and *ADH7*

(Hashibe *et al.* 2006; Hashibe *et al.* 2008; Akbari *et al.* 2009; Cui *et al.* 2009; Ding *et al.* 2009; Wu *et al.* 2012.a).

The histidine residue at *ADH1B* Arg48His (rs1229984) is associated with a decreased risk of OSCC in several populations, including Central European, Iranian and Japanese (Hashibe *et al.* 2006; Hashibe *et al.* 2008; Akbari *et al.* 2009; Cui *et al.* 2009). This polymorphism results in the "fast" oxidization of ethanol to acetaldehyde, which can be 100 times quicker than *ADH1B* Arg/Arg homozygotes (Hashibe *et al.* 2008). These results suggest that fast clearance of ethanol leads to a decreased risk of cancer. This would result in an increased rate of acetaldehyde production, which is perhaps counter-intuitive when *ALDH2* studies suggest that the build up of acetaldehyde increases risk of cancer, as discussed previously. However, normal activity of ALDH2 (Glu504Glu) may be sufficient to metabolize the resulting high levels of acetaldehyde. Indeed, the presence of the high-risk variants in both *ALDH2* and *ADH1B*, where ethanol is metabolized slowly and the resulting acetaldehyde cannot be oxidised, results in synergistic effects (Cui *et al.* 2009). These variants, together with smoking and alcohol consumption, have been shown in one study to result in a nearly 190 times increased risk of OSCC compared to individuals with low-risk *ALDH2* and *ADH1B* genetic variants and who abstain from drinking and smoking (Cui *et al.* 2009).

In individuals who process ethanol slowly (*ADH1B* Arg/Arg homozygotes), it is proposed that other mechanisms may come into effect to metabolize the compound. For example, ethanol may induce cytochrome P450 2E1 (CYP2E1) activation, which is another enzyme that converts ethanol into acetaldehyde, in *ADH1B* Arg48 homozygotes (Brooks *et al.* 2009.b). However, this reaction is also known to produce genotoxic oxygen radicals and lipid peroxidation products which may result in DNA damage, perhaps further contributing to cancer development (Brooks *et al.* 2009.b).

Variants in other alcohol dehydrogenase genes have also been associated with OSCC. The *ADH7* Gly92Ala (rs1573496) variant has been associated with a decreased risk of disease in European and Latin American populations, with a greater effect seen in drinkers, but the functional relevance of this polymorphism in unclear (Hashibe *et al.* 2008). *ADH1C* 350Val and 272Gln alleles have also been associated with an increased risk of upper aerodigestive tract cancers in Central Europeans but this is not consistent in other populations (Hashibe *et al.* 2006).

3.1.2 Apoptosis pathway

Apoptosis can be activated via three pathways; extrinsic, intrinsic or the granzyme B pathway. All three lead to activation of caspase-3 and/or-7, showing that caspases are essential in the apoptotic process. The extrinsic and intrinsic pathways are pictured below in Figure 3.1.

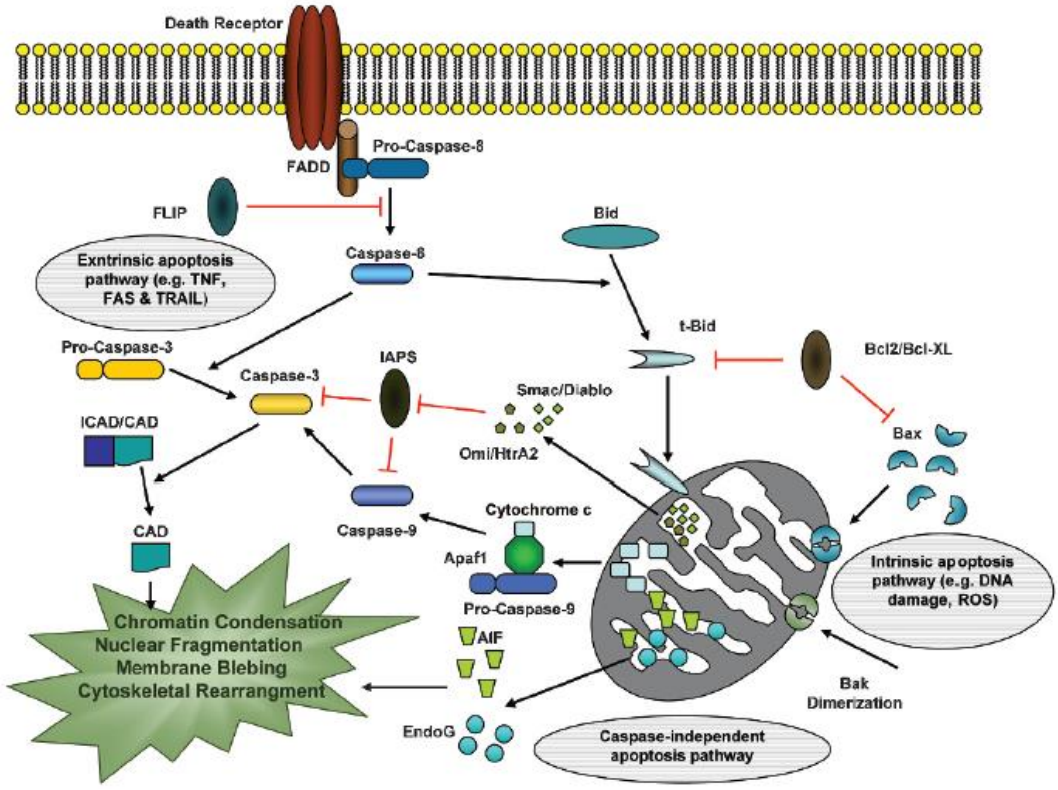


Figure 3.1: Apoptosis pathways
Includes the extrinsic, intrinsic and the caspase-independent apoptosis pathways (Ghavami *et al.* 2009).

The extrinsic pathway is initiated through the activation of cell surface death receptors including Fas (fibroblast associated antigen, also known as CD95) and TNF-R (tumour necrosis factor receptor) which contain a cytosolic death domain (DD) (reviewed in Ghavami *et al.* 2009). Following the binding of ligands (e.g. FASL) to these receptors, adaptor proteins such as FADD (Fas-associated protein with death domain) bind via their death domains to those of the receptor to form a death inducing signalling complex (DISC). Pro-caspase 8 is then recruited to the DISC and becomes activated through the cleavage of the pro-domain. Caspase-8 is able to activate downstream targets such as caspase-3 which leads to chromatin condensation, nuclear fragmentation, membrane blebbing and cytoskeletal rearrangement – the process of apoptosis.

3.1.2.1 Caspase-8

Several studies have found an association between caspase-8 mutations and cancer. Sun *et al.* (2007) identified a six-nucleotide deletion (-652 6Ndel) in the promoter region of *CASP8* which was associated with decreased risk of several cancers, including lung, oesophageal and breast cancers, in a Chinese population. Comparing 1,018 OSCC cases and 937 cancer-free controls, both the ins/del heterozygote and del/del homozygote were significantly associated with a decreased risk of OSCC with P=0.0278 (OR = 0.81; 95% CI = 0.67-0.98) and P=0.0082 (OR=0.57; 95% CI = 0.36-0.88), respectively.

Sun *et al*. proposed that the -652 deletion removes a stimulatory protein 1 (Sp1), a transcriptional activator, binding site, and showed that the deletion led to reduced *CASP8* transcription. As T lymphocytes are important in the management of cancer cells in the body, the effect of the *CASP-8* deletion in T cells was also studied. The deletion was shown to reduce activation-induced cell death (AICD) of T lymphocytes after cancer cell antigen stimulation. This implies that the immune surveillance would be more effective and, therefore, individuals with the deletion might be less susceptible to the development of cancer. The

authors also acknowledged that the deletion could be associated with an increased risk of cancer if the deletion occurred in cancer cells as they would be able to avoid apoptosis. However, Sun *et al.* pointed out that the majority of tumour cells are able to avoid apoptosis stimuli and that the extrinsic pathway of apoptosis is aberrant in malignant cells.

Conflicting results have been obtained for this polymorphism in other association studies both in OSCC and different cancer subtypes. In OSCC, a meta-analysis of two genome-wide association studies (2,961 cases and 3,400 controls) in a Han Chinese population did not find an association with a SNP in high LD (r^2 = 0.8) with the -652 6N deletion (Abnet *et al.* 2012). Additionally, the variant was not associated with OSCC in an northern Indian population (Umar *et al.* 2011). In other cancers, the deletion was associated with a decreased risk of squamous cell carcinoma of the head and neck in a non-Hispanic white population (Li *et al.* 2010.b) and also of bladder cancer in a Chinese population (Wang *et al.* 2009). No associations were identified in breast, colorectal and prostate cancer in an American multi-ethnic cohort study (Haiman *et al.* 2008) or in breast cancer in a European population (Frank *et al.* 2008). Two meta-analyses of all studies to date both found that the deletion was associated with a decreased overall risk of cancer (Yin *et al.* 2010; Zhang *et al.* 2012). However, even in these meta-analyses, results differed within cancer subtypes, with Yin *et al.* identifying an association with breast cancer whilst Zhang *et al.* did not. These results suggest that population and subtype-specific effects may exist.

Another *CASP8* polymorphism, D302H has also been associated with multiple cancers, including breast cancer (MacPherson *et al.* 2004; Cox *et al.* 2007). A meta-analysis showed that the variant affected susceptibility to the overall risk of cancers (Yin *et al.* 2010). No studies have analyzed the effect in OSCC, perhaps due the variant being non-polymorphic in Asian populations, in which the majority of these candidate gene case-control studies were done (Sun *et al.* 2007). The functional effect of this variant has yet to be determined, although

has been hypothesized to affect the autoprocessing of the procaspase-8 molecules or protein interactions with other molecules such as CFLAR (CASP8 and FADD-like apoptosis regulator) due to its location on the protein surface (MacPherson *et al.* 2004).

Interestingly, the meta-analysis of two Chinese OSCC GWAS, as discussed previously, did identify significant associations with variants at 2q33, a region containing *CASP8*, *ALS2CR12* and *TRAK2* (Abnet *et al.* 2012). The functional effects of these variants are unknown. The most significant association was observed for rs13016963, in the intron of *ALS2CR12*, ($P=7.63 \times 10^{-10}$), which was also in high LD with rs10931936 ($P=4.74 \times 10^{-9}$) located in an intron of *CASP8*. The identification of significant associations in the *CASP8* region in a hypothesis-free GWAS does provide extra support that this gene is involved in OSCC susceptibility, at least in a Chinese population. The inability to detect the association in each of the GWAS alone suggests that the effect size may be small, and thus, large sample numbers are needed.

In addition to the role of caspase-8 in the regulation of apoptosis, the protein is also involved in non-apoptotic pathways including monocyte differentiation, T cell activation, NF-κB activation and embryonic development (reviewed by Maelfait and Beyaert 2008). These functions may have a role in cancer development, but have not yet been explored.

3.1.2.2 FAS and FASL

Polymorphisms in *FAS* and its ligand, *FASL*, have been tested for association with cancer susceptibility, including OSCC. The two most studied variants are *FAS* -1377 G/A (rs2234767) and -670A/G (rs1800682). In a Han Chinese population, *FAS* -1377 AA genotype was associated with an increased risk of OSCC compared with GG genotypes (OR = 1.62, 95% CI = 1.14-2.30), as was *FAS* -670 GG compared to AA genotypes (OR = 1.57, 95% CI = 1.12-2.20) (Sun *et al.* 2004). Two meta-analyses of -1377 also suggest that the 'A' allele

increased overall cancer risk (Qiu *et al.* 2009; Zhang *et al.* 2009.a), although meaningful p-values were lacking from both studies. No association was observed for the -670 *FAS* variant in a cancer meta-analysis (Zhang *et al.* 2009.a).

Sun *et al.* (2004) also found *FASL* -844T/C (rs763110) variant to be associated with OSCC in the Han Chinese population, with the CC genotype increasing risk of disease compared to CT and TT genotypes (OR= 2.06, 95% CI = 1.64-2.59; P<0.001). This association was replicated in a meta-analysis of 19 studies from different types of cancers (Zhang *et al.* 2009.b). The presence of both *FAS*L -844CC and *FAS* -1377AA genotypes was associated with an even higher risk for OSCC (OR = 4.55, 95% CI = 2.75 – 7.78) (Sun *et al.* 2004).

All three of the *FAS* and *FASL* variants are located in potentially functional regions. The *FAS* -1377 A allele and the -670 G allele disrupt Sp1 and STAT1 transcription factor binding sites, respectively, which both lead to a decreased expression of *FAS* (Sun *et al.* 2004). *FASL* -844T/C is located in the promoter region within a C/EBP-β (CAAT/enhancer-binding protein β) transcription factor binding site, and the presence of the C allele at this position led to higher expression of *FASL* compared to the T allele (Wu *et al.* 2003).

In OSCC tissue compared to normal tissue, *FAS* expression is reduced, perhaps enabling tumour cells to evade apoptosis (Sun *et al.* 2004). In contrast, *FASL* expression is increased, which may increase the ability of tumour cells to counterattack the immune system by killing FAS-sensitive lymphocytes (Sun *et al.* 2004). However, more recently the role of *FAS* and *FASL* in cancer development has become more complex, with Chen *et al.* (2010) reporting that cancer cells require a constitutive activity of *FAS* for optimal growth. Here, independent knockdown of both *FAS* and *FASL* resulted in reduced tumour growth, implying that expression of both proteins is required in tumourigenesis. This growth-promoting role of FAS is in contrast to the traditional view that FAS

promotes cell death through its role as a death receptor. Therefore, FAS is proposed to have a secondary non-apoptotic role through activation of the JNK pathway which affects expression of *Egr1* and *Fos*, both of which are growth-promoting transcription factors.

3.1.3 Cyclooxygenase-2 (COX-2)

Cyclooxygenases are enzymes responsible for the rate-limiting step of the conversion of arachidonic acid to prostaglandins (PGs). Prostaglandins have a variety of roles, including in vasodilation, pain sensitivity, inflammation and cell proliferation (reviewed in Sobolewski *et al.* 2010).

COX-2 is an inducible isoform and can be activated by growth factors and inflammatory cytokines (reviewed in Cao and Prescott 2002 and Sobolewski et al. 2010). Over-expression of *COX-2* has been observed in several tumours including oesophageal squamous cell carcinoma (Ratnasinghe *et al.* 1999; Shamma *et al.* 2000). This is thought to lead to several processes in cancer development such as enhancing proliferation and survival of tumour cells, inhibiting apoptosis, tumour invasion and metastasis and angiogenesis (reviewed in Cao and Prescott 2002 and Sobolewski *et al.* 2010). Inhibition of *COX-2* expression in oesophageal cancer cell lines was shown to inhibit cell growth and induce apoptosis (Zimmermann *et al.* 1999). *COX-2* inhibition has, therefore, been a target for cancer therapies with non-specific COX inhibitors such as non-steroidal anti-inflammatory drugs (NSAIDs), and COX-2 specific inhibitors being developed for treatment and prevention of cancer.

Several studies have attempted to identify the functional variants of *COX-2* which affect the susceptibility to cancer. The polymorphism -765G>C (rs20417) has been association with an increased risk to OSCC in both northern Indian and Han Chinese populations (Zhang *et al.* 2005; Upadhyay *et al.* 2009). This mutation is located within the *COX-2* promoter and the ‘C’ allele is proposed to disrupt a binding site for the transcriptional activator SP1 (stimulatory protein 1)

and to create an E2F transcription factor binding site (Papafili *et al.* 2002; Zhang *et al.* 2005). Functional analyses of the effect of different alleles at this variant have produced conflicting results with the -765C allele resulting in a decreased expression of *COX-2* in human lung fibroblasts (Papafili *et al.* 2002), but producing a >10-fold increase in prostaglandin levels in a monocyte culture (for -765CC genotype compared to GG) (Szczeklik *et al.* 2004). Zhang *et al.* (2005) suggest that this is plausible due to the creation of the E2F transcription factor binding site.

The variant -1195G>A (rs689466) in the *COX-2* promoter has also been associated with an increased risk to OSCC in a Han Chinese population (Zhang *et al.* 2005) but not in a northern Indian population (Upadhyay *et al.* 2009). The 'A' allele creates a c-MYB (a transcription factor) binding site which results in an increased expression of *COX-2* (Zhang *et al.* 2005). c-MYB is involved in the regulation of cell division, differentiation and cell survival and is known to activate *COX-2* (Ramsay *et al.* 2003), which supports the role of the -1195G>A polymorphism in the susceptibility to cancer.

3.1.4 *O*^{6}-methylguanine-DNA methyltransferase (MGMT)

O^6-methylguanine-DNA methyltransferase (MGMT) is involved in DNA repair and catalyzes the irreversible transfer of the methyl group from *O*^{6}-methylguanine (*O*^{6}-MeG) adducts to its own cysteine molecule (reviewed in Kaina *et al.* 2007). *O*^{6}-MeG is formed due to the alkylation of DNA by alkylating agents, which may be present in the environment, such as in tobacco smoke or food, or as a product of cellular metabolic processes. This alkylation causes a transition mutation from G to A, resulting in a A:T pairing instead of G:C. If this point mutation occurs in genes involved in the regulation of cell proliferation, such as *TP53* or *KRAS*, then this may initiate cancer development (reviewed in Esteller and Herman 2004). Mismatch repair mechanisms can repair the thymine residue, but unless *O*^{6}-MeG is repaired by MGMT, thymine continues to be inserted on the opposite DNA strand (reviewed in Kaina *et al.* 2007). If

MGMT is not present, DNA-double strand breaks accumulate due to the collapse of replication forks, leading to apoptosis (reviewed in Kaina *et al.* 2007). Therefore, the presence of MGMT is essential to convert O^6-MeG to its non-mutagenic form, to allow the normal functioning of the cell.

Several polymorphisms in *MGMT* have been associated with susceptibility to cancer, although studies have produced conflicting results. Phenylalanine at *MGMT* Leu84Phe (rs12917) is associated with increased risk for upper aerodigestive tract (UADT) cancers in Central and Eastern European countries (Hall *et al.* 2007). However, the same SNP was associated with a decreased risk of head and neck cancer in an American Caucasian population (Huang *et al.* 2005). A meta-analysis of all cancer studies published found *MGMT* 84Phe to be associated with an increased risk of cancer (Zhong *et al.* 2010). The functional significance of Leu84Phe is unclear. One suggestion if that the polymorphism might affect the ability of *MGMT* to inhibit endoplasmic reticulum-mediated cell proliferation (Zhong *et al.* 2010).

3.2 Candidate gene association studies in the South African populations

As discussed, variants in genes involved in apoptosis, DNA repair and alcohol metabolism have been associated with OSCC and other cancers. However, results across populations and in different cancer sub-types are not always consistent. These variants were investigated to determine whether they were associated with the susceptibility to OSCC in South African populations in my paper entitled “Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa” which is presented in journal format overleaf. Details of the Materials and Methods are given in Chapter 2. My contribution to this paper has been to perform all of the genotyping and the statistical analysis. I also wrote the paper, together with C. Mathew.

Carcinogenesis vol.32 no.12 pp.1855–1861, 2011
doi:10.1093/carcin/bgr211
Advance Access publication September 16, 2011

Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa

Hannah Bye[1], Natalie J.Prescott[1], Marco Matejcic[2], Elizabeth Rose[3], Cathryn M.Lewis[1], M.Iqbal Parker[2] and Christopher G.Mathew[1,*]

[1]Department of Medical and Molecular Genetics, King's College London, King's Health Partners, Guy's Hospital, London SE1 9RT, UK, [2]International Centre for Genetic Engineering and Biotechnology and Division of Medical Biochemistry, University of Cape Town, Cape Town, South Africa and [3]Medical Research Council, Tygerberg, South Africa

*To whom correspondence should be addressed. Tel: +44 (0) 20 7188 3713; Fax: +44 (0) 207 188 2585;
Email: christopher.mathew@kcl.ac.uk

Genetic variants in multiple cellular pathways have been associated with an altered risk of oesophageal cancer. In this study, eight genes previously associated with an altered risk of oesophageal squamous cell carcinoma (OSCC) in European or Asian populations were investigated in two South African populations. We genotyped 12 single-nucleotide polymorphisms and one insertion/deletion variant in 1463 individuals from the Black and Mixed Ancestry populations. No polymorphisms were associated with OSCC in the Black population. In the Mixed Ancestry population, *ALDH2* +82 G > A (rs886205) was significantly associated with a reduced risk of OSCC (odds ratio = 0.70, 95% confidence interval = 0.55–0.89; $P = 0.0038$). Several other polymorphisms showed a suggestive association ($P < 0.05$), including *ADH1B* Arg48His (rs1229984), *COX-2* −1195G > A (rs689466), *CASP8* Asp302His (rs1045485) and *MGMT* Leu84Phe (rs12917). Haplotype analysis indicated that the *FAS* polymorphisms −670 A > G (rs1800682) and −1377 G > A (rs2234767) were both associated with OSCC in the Mixed Ancestry population ($P = 0.006$ and $P = 0.004$, respectively), as well as the *CASP8* (−652 6Ndel:302His) haplotype ($P = 0.0013$). This study indicates several instances of population-specific differences in the genetic etiology of OSCC between these two South African populations and between them and other high-risk populations, which may reflect differences in their ancestry and environmental exposures.

Introduction

Oesophageal cancer is the eighth most common cancer in the world and is responsible for >300 000 deaths a year (1). The disease has a very poor prognosis with a 5 years survival rate of <10% (2). Two main subtypes exist, squamous cell carcinoma and adenocarcinoma, which are etiologically unrelated. Oesophageal squamous cell carcinoma (OSCC) is the predominant form in developing countries (3). High-risk regions have been identified in China, Japan, Iran and southern Africa. In the Eastern Cape Province of South Africa, oesophageal cancer is the most common malignancy in Black males and the second most common in Black females, with an incidence of 32.7 and 20.2 cases per 100 000 people, respectively (4). Alcohol and tobacco are implicated in the majority of cases in the western world (1). In South Africa, additional risk factors include nutritional deficiencies, consumption of maize contaminated with the *Fusarium* fungus and human papilloma virus infection [reviewed in Hendricks *et al.* (2)].

Most genetic studies in OSCC have focused on candidate genes involved in alcohol metabolism, detoxification of carcinogens, DNA repair, apoptosis and cell proliferation [reviewed in Lao-Sirieix *et al.* (5)]. However, the results have not always been consistent, particularly across different populations. This may reflect differences in the prevalence of susceptibility variants between populations, differences in environmental exposures or technical issues such as small sample sizes which are not well powered to detect modest genetic effects. Genome-wide association studies (GWAS) in Japanese and Chinese populations have detected association of genetic variants in *ADH1B*, *ALDH2*, *PLCE1* and *C20orf54* with OSCC (6–8). Recently, a GWAS in upper aerodigestive cancers including OSCC in European populations reported associations in *ADH7*, the *ALDH2* locus and a novel association in the DNA repair gene *HEL308* (9). Our previous studies in the South African population have detected association of genetic variants in several genes with OSCC, including *GSTP1* (10), *CYP2E1* (11), *SULT1A1* and *CYP3A5* (12). In this study, we have sought to obtain a clearer understanding of genetic and environmental factors contributing to the pathogenesis of OSCC in an expanded cohort from the Black and Mixed Ancestry populations of South Africa by investigation of 12 single-nucleotide polymorphisms (SNPs) and one insertion/deletion variant from eight genes with previous robust evidence of association with OSCC in other populations.

Abbreviations: CI, confidence interval; OR, odds ratio; GWAS, genome-wide association study; LD, linkage disequilibrium; OR, odds ratio; OSCC, oesophageal squamous cell carcinoma; PCR, polymerase chain reaction.

Materials and methods

Study subjects

A total of 1463 individuals were recruited from the Black and Mixed Ancestry populations of South Africa. The Black subjects were mainly Xhosa-speakers from the Eastern or Western Cape of South Africa, who are one of the major populations originating from the Bantu-speaking peoples of Southern Africa. The Mixed Ancestry subjects were from the Western Cape. This population (also referred to in the literature and self-reported as the 'coloured' population of South Africa) is an admixed population with major ancestral components from the indigenous Khoisan, Bantu-speaking Africans, Europeans and Asians (13). The study consisted of 358 OSCC patients and 477 controls from the Black population and 201 OSCC patients and 427 controls from the Mixed Ancestry population. All patients were recruited between March 2000 and September 2010 at Groote Schuur Hospital (GSH), Cape Town, South Africa, with histologically confirmed primary invasive OSCC. Control samples were recruited from the same populations as the patients and from the same geographical area, age group, gender and ethnic group. Data on alcohol and tobacco use were available for OSCC cases. Smoking status was subdivided into those who were current smokers, former smokers or never-smokers. Drinkers were defined as subjects who consumed alcohol at least once in every week. Demographic and exposure data are given in Table I. Whole blood samples were collected with informed consent from all subjects and DNA was extracted at the University of Cape Town. Ethical approval for the study was obtained from the joint University of Cape Town/GSH Research Ethics Committee.

Candidate genes

Polymorphisms were selected for genotyping following a literature review of published genetic association studies involving OSCC and other head-and-neck squamous cell carcinomas (Table II). A total of 12 SNPs and one insertion/deletion variant were genotyped in cases and controls: *ADH1B* Arg48His (rs1229984); *ADH7* Gly92Ala (rs1573496); *ALDH2* Glu504Lys (rs671), +82 A > G (rs886205) and −261 C > T (rs441); *FAS* −670 G > A (rs1800682) and −1377 G > A (rs2234767); *FASL* −844 T > C (rs763110); *COX-2* −765 G > C (rs20417) and −1195 A > G (rs689466); *MGMT* Leu84Phe (rs12917); *CASP8* Asp302His (rs1045485) and *CASP8* −652 6N ins/del (rs3834129). It is important to note that the nomenclature for these variants is based on older versions of the human genome and nucleotide annotation. However, for simplicity and to allow consistency with previous literature in the field, we have maintained this 'common nomenclature' throughout. A summary of the updated nomenclature for these variants using the more recent genome annotation (NCBI36) adopted by the Human Genome Variation Society (www.HGVS.org) is provided in Supplementary Table I, available at *Carcinogenesis* Online.

SNP genotyping

All samples were genotyped using the TaqMan 5′ exonuclease assay with primers and probes designed and synthesized by Applied Biosystems (Carlsbad, CA) (19).

SNPs in the alcohol and aldehyde dehydrogenase genes (with the exception of *ALDH2* −261 T > C) were genotyped using validated TaqMan drug metabolism genotyping assays (Applied Biosystems). SNPs in *MGMT*, *FAS* and *COX-2* −1195 A > G were genotyped using validated TaqMan SNP-genotyping assays (Applied Biosystems). Custom TaqMan assays were designed for *COX-2* −765G > C [forward primer, CCCCCTCCTTGTTTCTTGGAA; reverse primer, TGCTTAG GACCAGTATTATGAGGAGAA; reporter ACCTTTCCC(G/C)CCTCTC], *CASP8* Asp302His [forward primer, ACCACGACCTTTGAAGAGCTT; reverse primer, TCCATGAGTTGGTAGATTTTCAAAATCTCA; reporter CCCCAC(G/C)]ATGACTG] and *ALDH2* −261 G > C (forward primer, AGCCTGGGTGCCAGAGAGA; reverse primer, CCTGACAGCATTCACTTAGAACAAC; reporter 1, CTCGGCCTCAAAA; reporter 2, ACTCGGTCTCAAAAA). Reactions were carried out in 2.5 µl volumes in 96-well plates. Each reaction contained 20 ng DNA, Absolute QPCR ROX mix (Abgene, Epsom, UK) and SNP assay mix (Applied Biosystems) according to assay instructions and were performed on a PTC-0225 DNA Engine (MJ Research, Waltham, MA). Fluorescent levels at the polymerase chain reaction (PCR) end-point were determined using a 7900HT Fast Real-Time PCR system (Applied Biosystems) and genotypes assigned using SDS 2.2.2 software (Applied Biosystems).

Insertion/deletion genotyping

Primers for the *CASP8* −652 6N ins/del were previously designed by Sun *et al.* (25). Briefly, the 5 µl PCR reaction contained 1× PCR mastermix (Promega, Madison, WI), 0.4 µM of each primer (Sigma, Dorset, UK), 10 ng of DNA and was performed on a thermocycler as above. The PCR products were separated by capillary electrophoresis on an ABI3730xl DNA Analyzer (Applied Biosystems) and sized using GeneMapper software (Applied Biosystems).

Statistical analysis

Pearson's chi-squared (χ^2) test was used to determine deviations from the Hardy–Weinberg equilibrium; all genotype frequencies were in Hardy–Weinberg equilibrium in both populations, with the exception of *FAS* -670 G > A in OSCC cases from the Black population ($P = 0.027$). Genotype and allele frequencies were calculated for cases and controls and compared using the Pearson's chi-squared (χ^2) test to test for association with OSCC. A *P*-value of <0.0042 (0.05/12) was used as a significance threshold for the association test to allow for multiple testing of the 12 variants present in these populations based on the Bonferroni principle. No additional correction was applied for the two populations tested. Genotypic and allelic odds ratios (ORs) with 95% confidence intervals (CIs) were calculated using the common homozygous genotype or common allele as the reference. Haplotype analysis and determination of linkage disequilibrium (LD) between variants in the same gene were performed using UNPHASED (26). SNPs that were suggestive of an allelic association (uncorrected $P < 0.05$) were further investigated for the effect of alcohol and tobacco by stratifying cases based on smoking and drinking status.

Table I. Characteristics of OSCC cases in the South African Black and Mixed Ancestry patients

	Black population	Mixed Ancestry population
Controls	*n* = 477	*n* = 427
Cases	*n* = 358	*n* = 201
Summary statistics—cases		
Age, mean years (SD)	59.8 (11.3)	60.5 (10.6)
Sex, *n* (%)		
Male	182 (50.8)	131 (65.2)
Female	176 (49.2)	70 (34.8)
Smoking status, *n* (%)		
Current smoker	100 (27.9)	131 (65.2)
Former smoker	128 (35.8)	58 (28.9)
Never smoker	130 (36.3)	10 (5.0)
Unknown	0	2 (1)
Alcohol consumption, *n* (%)		
Drinker	228 (63.7)	163 (81.1)
Non-drinker	128 (35.8)	37 (18.4)
Unknown	2 (0.6)	1 (0.5)

Results

Case–control analysis

The results for the case–control analysis in the two South African populations are shown in Table III and full genotype counts in Supplementary Table II, available at *Carcinogenesis* Online. None of the 13 variants tested were associated with OSCC in the Black South African population. The *ADH1B* 48His and *ALDH2* 504Lys alleles were absent in this population, and the *ADH7* 92Ala allele was extremely rare (only one allele observed).

In the Mixed Ancestry population, the SNP *ALDH2* +82G > A (rs886205) showed association with OSCC ($P = 0.0038$), which remained significant after accounting for multiple testing. The +82A allele had a frequency of 40.2% in cases and 48.9% in controls giving an allelic OR of 0.70 (95% CI = 0.55–0.89) and was thus associated with a reduced risk of OSCC. Suggestive associations were observed in the Mixed Ancestry population for several other polymorphisms ($P < 0.05$): *ADH1B* Arg48His ($P = 0.009$), *COX-2* −1195 A > G ($P = 0.014$), *CASP8* Asp302His ($P = 0.040$) and *MGMT* Leu84Phe ($P = 0.023$). However, these associations did not meet the required threshold for multiple testing. Variants not showing significant associations were *COX-2* −765 G > C, *ALDH2* −261 T > C, *FAS* −670 G > A,

Table II. Summary of association results in published studies

Gene	Common variant name[a]	dbSNP ID	Genetic model	OR (95% CI)	*P*-value	Cancer site	Population	Ref
ALDH2	Glu504Lys (G > A)	rs671	A versus G	1.67 (1.58–1.76)	3.27×10^{-24}	OSCC	Japanese	(7)
ALDH2	+82 A > G	rs886205	GG versus AA	4.14 (2.03–8.46)	<0.0001	OSCC	European	(14)
ALDH2	−261 C > T	rs441	CC versus TT	3.85 (1.78–8.36)	<0.0001	OSCC	European	(14)
ADH1B	Arg48His (G > A)	rs1229984	A versus G	1.79 (1.69–1.88)	7.75×10^{-24}	OSCC	Japanese	(7)
			GG + GA versus AA	0.34 (0.20–0.56)	—	OSCC	European and Latin American	(15)
ADH7	Gly92Ala	rs1573496	C versus G	0.45 (0.32–0.64)	—	OSCC	European and Latin American	(15)
			GG versus CC	0.32 (0.13–0.82)	—	Head and neck SCC	American Caucasian	(16)
FAS	−670 A > G	rs1800682	GG versus AA	1.57 (1.12–2.20)	<0.001	OSCC	Chinese	(17)
FAS	−1377 G > A	rs2234767	AA versus GG	1.62 (1.14–2.30)	<0.001	OSCC	Chinese	(17)
FASL	−844 T > C	rs763110	CC versus TT	1.72 (1.12–2.64)	<0.001	OSCC	Chinese	(17)
CASP8	Asp302His (G > C)	rs1045485	CC versus GG	0.81 (0.71–0.93)	—	Cancer meta-analysis	—	(18)
CASP8	−652 6N ins/del	rs3834129	del/del versus ins/ins	0.57 (0.36–0.88)	0.0082	Oesophageal cancer	Chinese	(19)
COX-2	−765 G > C	rs20417	GC versus GG	2.24 (1.59–3.16)	<0.0001	OSCC	Chinese	(20)
			C versus G	1.32 (0.92–1.88)	—	OSCC	Indian	(21)
COX-2	−1195 G > A	rs689466	A versus G	1.34 (1.08–1.68)	0.008	OSCC	Chinese	(20)
			A versus G	1.23 (0.80–1.87)	—	OSCC	Indian	(21)
MGMT	Leu84Phe	rs12917	FF versus LL	3.27 (1.43–7.52)	—	OSCC	European	(22)
			LF + FF versus LL	0.71 (0.51–0.98)	—	Head and neck SCC	American	(23)
			TT versus CC	1.24 (1.02–1.51)	0.035	Cancer meta-analysis	—	(24)

[a]Variant names correspond to common names used in majority of previous publications. Updated variant names based on more recent annotations of the human genome are provided in Supplementary Table I, available at *Carcinogenesis* Online.

FAS −1377 G > A, *FASL* −844T > C and *CASP8* −652 6N ins/del. The *ALDH2* 504Lys allele was absent in Mixed Ancestry subjects, as in the Black population.

Haplotype analysis

Multiple polymorphisms were genotyped in *COX-2*, *ALDH2*, *FAS* and *CASP8* genes, enabling haplotype analysis to be performed and LD to be determined (Table IV). A low level of LD was observed between variants in the different genes in both the Black and Mixed Ancestry population. The LD coefficients (r^2) for pairs of SNPs in *COX-2*, *ALDH2*, *FAS* and *CASP8* were 0.058, 0.058, 0.023 and 0.021 in the Black population controls and 0.087, 0.228 and 0.153 and 0.001 in the Mixed Ancestry population controls, respectively. Similar values were observed in OSCC cases.

In the Black population, there were no significant haplotype effects. In the Mixed Ancestry population, statistically significant haplotype associations were observed for variants in *ALDH2* ($P = 0.0028$), *FAS* ($P = 0.0031$) and *CASP8* ($P = 0.004$). However, the haplotype result observed for *ALDH2* is entirely due to the association of the *ALDH2* +82 allele observed in the single SNP analysis, with no increase in significance achieved by inclusion of the −261 variant. Haplotypes at the *FAS* gene locus were significantly associated with a reduced risk of OSCC in the Mixed Ancestry population (overall $P = 0.003$), with both −1377A and −670A alleles contributing independently to disease risk, consistent with the low level of LD between them. For *CASP8*, the haplotype −652 6Ndel:302His was significantly associated with an increased risk of OSCC ($P = 0.001$, OR = 2.37; 95% CI = 1.39–4.04), whereas the individual variants were not.

Alcohol and smoking analysis

Polymorphisms that showed a statistically significant ($P < 0.0042$) or suggestive ($P < 0.05$) association with OSCC in the case–control analysis were investigated further for gene–environment interactions. These were *ADH1B* Arg48His, *COX-2* −1195 A > G, *CASP8* Asp302His, *ALDH2* +82 G > A and *MGMT* Leu84Phe in the Mixed Ancestry population only. In this population, current, former and never-smokers account for 65.2% ($n = 131$), 28.9% (58) and 5% (10) of OSCC patients, respectively (Table I). In view of the low numbers of never-smokers, analyses were carried out only for current and former smokers, with each group being compared with controls (Table V). All SNPs analyzed showed nominal evidence of association with OSCC in current smokers ($P < 0.05$) and no associations in the smaller group of former smokers. The most significant association in current smokers compared with controls was for *MGMT* Leu84Phe ($P = 0.003$) and was more significant than the association seen in the initial case–control test for all cases combined ($P = 0.023$), with a concomitant increase in the disease risk (OR all cases 1.41 and OR current smokers 1.65).

Alcohol drinkers represent 81.1% of the Mixed Ancestry patients (Table II). Comparing drinkers to controls, all SNPs analyzed showed at least nominal evidence of association ($P < 0.05$), with the two most significant findings at *ALDH2* +82 G > A and *COX-2* −1195 A > G ($P = 0.003$ and $P = 0.004$, respectively) (Table VI), achieving greater significance in this stratified analysis compared with the initial analysis of all cases combined. Analysis of drinkers versus non-drinkers showed no significant differences between the groups as did non-drinkers versus controls, but the number of non-drinkers was small.

Discussion

In this study, we tested 13 sequence variants for association with OSCC in eight genes involved in several candidate molecular pathways, including alcohol metabolism, apoptosis, cell proliferation and DNA repair. No associations were observed in the Black South African population, whereas several significant or suggestive associations were detected in the Mixed Ancestry population. Possible explanations for the differences between these two South African populations are discussed below.

In the Mixed Ancestry population, one SNP, *ALDH2* +82 A > G (rs886205), was significantly associated with OSCC after accounting for multiple testing. Another SNP upstream of *ALDH2*, −261C > T, was not associated with OSCC in this population and the 504Lys allele was absent. ALDH2 metabolizes acetaldehyde into acetate and substitution of glutamic acid by lysine at amino acid position 504 results in a catalytically inactive subunit (27). The 504Lys allele is thought to be almost unique to Asian populations (14), and this study confirms its absence in two South African populations. The +82A allele was associated with a reduced risk of OSCC (OR = 0.70) in the Mixed Ancestry population, thus replicating the association of this SNP with OSCC previously observed in Central European populations. It should be noted that in Europeans, the +82G is the minor allele, which has an increased frequency in OSCC patients and is therefore reported as increasing cancer risk (28). Despite the difference in allele frequency between these two populations, these two observations are consistent in that the same allele (G) is increased in frequency in both OSCC populations. This SNP, located 360 bp upstream of the ATG initiation codon for *ALDH2*, is within or adjacent to known or predicted binding sites for multiple transcription factors (17,29,30). Analysis of the effect of this variant on transcriptional activity has produced conflicting results; the +82G allele has been shown to be more active than the A allele in hepatoma cells (29), but the opposite was observed in human peripheral blood leukocytes when analyzing the basal level of transcription (17). However, the latter study showed that higher levels of expression from the G allele

Table III. Association of polymorphisms with OSCC in the South African Black and Mixed Ancestry populations

Gene	SNP	Alleles (major + reference/minor)	Black population				Mixed Ancestry population			
			MAF		OR (95% CI)	*P*-value	MAF		OR (95% CI)	*P*-value*
			Cases	Controls			Cases	Controls		
ADH1B	Arg48His	G/A	0	0.000	—	—	0.054	0.098	0.52 (0.32–0.86)	0.009
ADH7	Gly92Ala	C/G	0	0.001	—	—	0.014	0.02	0.67 (0.22–2.01)	0.471
ALDH2	+82 G > A	G/A	0.247	0.252	0.98 (0.78–1.23)	0.835	0.402	0.489	0.70 (0.55–0.89)	0.004
ALDH2	−261 T > C	T/C	0.154	0.145	1.07 (0.81–1.42)	0.611	0.18	0.194	0.92 (0.67–1.25)	0.587
COX-2	−765 G > C	G/C	0.471	0.513	0.85 (0.69–1.03)	0.096	0.376	0.321	1.28 (0.99–1.64)	0.059
COX-2	−1195 A > G	A/G	0.064	0.053	1.22 (0.80–1.86)	0.343	0.103	0.155	0.63 (0.43–0.91)	0.014
MGMT	Leu84Phe	C/T	0.189	0.195	0.96 (0.75–1.24)	0.770	0.222	0.168	1.41 (1.05–1.91)	0.023
CASP8	Asp302His	G/C	0.154	0.152	1.02 (0.77–1.34)	1.000	0.169	0.126	1.42 (1.01–1.98)	0.040
CASP8	−652 6N ins/del	Ins/Del	0.518	0.502	1.06 (0.87–1.30)	0.530	0.385	0.386	0.99 (0.77–1.27)	1.000
FAS	−670 G > A	G/A	0.219	0.225	0.96 (0.76–1.22)	0.750	0.356	0.406	0.81 (0.63–1.04)	0.097
FAS	−1377 G > A	G/A	0.096	0.072	1.36 (0.95–1.94)	0.092	0.139	0.183	0.72 (0.52–1.01)	0.058
FASL	−844 T > C	T/C	0.192	0.189	1.02 (0.79–1.31)	1.000	0.416	0.386	1.13 (0.89–1.45)	0.323

*P-value corrected to three decimal places.

Table IV. Haplotype analysis for polymorphisms in *COX-2*, *FAS*, *ALDH2* and *CASP8* in the South African Black and Mixed Ancestry populations

	Black population					Mixed Ancestry population				
	Cases (%)	Controls (%)	OR (95% CI)	*P*-value	*P*-value for overall haplotype association	Cases (%)	Controls (%)	OR (95% CI)	*P*-value	*P*-value for overall haplotype association
COX-2					0.254					0.019
−765G −1195A	318 (46.4)	399 (43.7)	1.15 (0.94–1.41)	0.184		196 (51.9)	436 (52.3)	Ref	Ref	
−765C −1195A	324 (47.2)	467 (51.1)	Ref	Ref		143 (37.8)	268 (32.1)	1.19 (0.92–1.55)	0.204	
−765G −1195G	44 (6.4)	48 (5.3)	1.32 (0.86–2.04)	0.208		39 (10.3)	130 (15.6)	0.67 (0.45–0.99)	0.041	
FAS					0.238					0.003
−670G −1377G	464 (68.6)	637 (70.0)	Ref	Ref		195.6 (51.2)	339 (41.0)	Ref	Ref	
−670A −1377G	147 (21.8)	207 (22.8)	0.97 (0.76–1.24)	0.838		133.4 (34.9)	336 (40.7)	0.69 (0.53–0.90)	0.006	
−670G −1377A	65 (9.6)	66 (7.3)	1.35 (0.94–1.94)	0.104		51.38 (13.5)	151 (18.3)	0.59 (0.41–0.85)	0.004	
ALDH2					0.936					0.003
+82A −261T	166 (24.7)	231 (25.2)	0.98 (0.77–1.24)	0.872		155.3 (40.0)	405 (48.7)	Ref	Ref	
+82G −261T	403 (6.0)	550 (60.0)	Ref	Ref		162.7 (41.9)	266 (32.0)	1.60 (1.22–2.09)	0.001	
+82G −261C	103 (15.3)	135 (14.7)	1.04 (0.78–1.39)	0.782		68.32 (17.6)	161 (19.4)	1.11 (0.79–1.55)	0.559	
CASP8					0.406					0.004
−652 6N Ins 302D	256 (39.7)	361.2 (39.9)	0.99 (0.79–1.26)	0.959		200.2 (54.4)	436 (53.0)	Ref	Ref	
−652 6N Del 302D	290 (45.0)	406.8 (44.9)	Ref	Ref		102.8 (27.9)	281 (34.2)	0.80 (0.59–1.07)	0.126	
−652 6N Ins 302H	52.04 (8.1)	92.78 (10.2)	0.79 (0.53–1.18)	0.238		25.8 (7.0)	69 (8.4)	0.81 (0.48–1.39)	0.448	
−652 6N Del 302H	45.96 (7.1)	45.22 (5.0)	1.43 (0.83–2.44)	0.199		39.2 (10.7)	36 (4.4)	2.37 (1.39–4.04)	0.001	

could be induced by acetaldehyde or ethanol, suggesting that this SNP may contribute to interindividual or allelic differences in acetaldehyde elimination (17). The association of *ALDH2* +82A > G in the South African Mixed Ancestry population is consistent with many other studies which have found associations of other variants at this gene with oesophageal cancer, including patients of European, Chinese and Japanese descent (8,9, 28,31–33). Interestingly, this is a region of extended LD in European populations, which contains other plausible candidate genes (9).

The SNPs *ADH1B* Arg48His, *COX-2* −1195A > G, *CASP8* Asp302His and *MGMT* Leu84Phe showed some evidence of association with OSCC in the Mixed Ancestry population with $P < 0.05$ but did not survive the Bonferroni threshold which was set for multiple testing. However, given that this threshold is somewhat conservative and that there is prior evidence for the association of these variants with OSCC or head and neck squamous cell carcinoma in other populations, we have referred to these as suggestive associations with OSCC which require future follow-up in an expanded sample.

The associations of haplotypes at the *FAS* gene locus which contained either the −1377A or −670A alleles with a reduced risk of OSCC in the Mixed Ancestry population are only partially consistent with data from the Han Chinese population, in which −670A was associated with a reduced risk of OSCC but −1377A had the opposite effect (34). The *FAS* protein and its ligand, *FASL*, play a key role in the induction of apoptosis, and *FAS* expression is found to be reduced in OSCC tissues (35,36). However, a recent study by Chen *et al.* (37) showed that constitutive expression of *FAS* is required for optimal growth of tumors and complete loss of *FAS* is rarely observed. Functional studies have shown that *FAS* −1377A and −670G disrupt the Sp1 and STAT1 transcription factor-binding sites, respectively, which both lead to reduced expression of *FAS* (15,38,39). Taken together, these data suggest that the *FAS* −1377A and −670G alleles would both be associated with an increased risk of cancer, provided that *FAS* is constitutively expressed, but population-specific effects may occur.

In our study, neither of the *CASP8* variants, Asp302His and -652 6N ins/del, were significantly associated with OSCC. However, in the haplotype analysis, the *CASP8* −652 6Ndel:302His haplotype was significantly associated with an increased risk OSCC. Previous independent association studies of *CASP8* variants with cancer have produced somewhat conflicting results. The −652 6N deletion allele has been associated with reduced susceptibility to multiple cancers, including oesophageal cancer, in the Chinese population (25), but this was not replicated for breast, colorectal and prostate cancer in several other populations (40). A meta-analysis of 55 studies reported a reduced overall risk of cancer for the −652 6N deletion and 302His alleles, although stratified analysis suggested a reduced risk for estrogen-related cancers but an increased risk for brain tumors (18). It is possible that population-specific or cancer subtype-specific effects may occur, and results may also be influenced by differing environmental triggers.

Gene–environment interactions are known to exist in susceptibility to OSCC, particularly those involving an interaction between alcohol metabolism genes and alcohol intake or smoking status. The risk of disease conferred by the *ALDH2* 504Lys allele, for example, is related to the amount of alcohol consumed (8, 41), but association studies in non-drinkers have reported conflicting results (8, 32, 41). Smoking status also appears to interact with this variant, with smokers having a greater risk of disease than non-smokers (8). The association observed in this study for *ALDH2* +82A > G in the South African Mixed Ancestry population was somewhat stronger when the data were stratified to include only those cases that drink. This follows a similar trend to that observed by Hashibe *et al.* (28), who found an increased disease risk for A/G heterozygotes and G/G homozygotes in both light and medium/heavy drinkers in a European population. Gene–environmental interactions involving *ADH1B* His48Arg have also been conflicting. Cui *et al.* (8) found no interaction of this SNP with drinking and smoking status in the Japanese population. This is in contrast to a study of upper aerodigestive cancers,

Table V. Analysis of the effect of smoking for polymorphisms that show a significant or suggestive association with OSCC in the Mixed Ancestry population

Gene	Current smokers (n = 131) versus controls (n = 427)			Former smokers (n = 58) versus controls (n = 427)		
	MAF: controls/ current smokers	OR (95% CI)	P-value	MAF: controls/ former smokers	OR (95% CI)	P-value
ADH1B Arg48His	0.098/0.047	0.45 (0.24–0.84)	0.011	0.098/0.096	0.98 (0.51–1.91)	1.000
COX-2 −1195 A > G	0.155/0.100	0.61 (0.38–0.95)	0.029	0.155/0.129	0.81 (0.46–1.44)	0.468
CASP8 Asp302His	0.126/0.180	1.53 (1.04–2.24)	0.028	0.126/0.179	1.51 (0.90–2.56)	0.119
ALDH2 +82 G > A	0.489/0.418	0.75 (0.56–0.99)	0.045	0.489/0.407	0.72 (0.48–1.06)	0.092
MGMT Leu84Phe	0.168/0.250	1.65 (1.18–2.32)	0.003	0.168/0.186	1.14 (0.69–1.87)	0.615

Table VI. Analysis of the effect of alcohol consumption for polymorphisms that show a significant or suggestive association with OSCC in the Mixed Ancestry population

Gene	Case-only analysis: drinkers (n = 163) versus non-drinkers (n = 37)			Drinkers (n = 163) versus control (n = 427)			Non-drinkers (n = 37) versus controls (n = 427)		
	MAF: non-drinkers/ drinkers	OR (95% CI)	P-value	MAF: controls/ drinkers	OR (95% CI)	P-value	MAF: controls/ non-drinkers	OR (95% CI)	P-value
ADH1B Arg48His	0.042/0.057	1.38 (0.40–4.82)	0.612	0.098/0.057	0.55 (0.33–0.94)	0.026	0.098/0.042	0.40 (0.12–1.30)	0.116
COX-2 −1195 A > G	0.162/0.090	0.51 (0.25–1.06)	0.066	0.155/0.090	0.54 (0.35–0.83)	0.004	0.155/0.162	1.05 (0.55–2.01)	0.873
CASP8 Asp302His	0.111/0.184	1.80 (0.82–3.97)	0.139	0.126/0.184	1.57 (1.10–2.23)	0.012	0.126/0.111	0.87 (0.41–1.87)	0.721
ALDH2 +82 G > A	0.432/0.394	0.85 (0.51–1.42)	0.541	0.489/0.394	0.68 (0.52–0.88)	0.003	0.489/0.432	0.79 (0.49–1.28)	0.347
MGMT Leu84Phe	0.176/0.234	1.43 (0.75–2.76)	0.277	0.168/0.234	1.52 (1.10–2.08)	0.010	0.168/0.176	1.06 (0.57–1.97)	0.863

including OSCC, in Central European countries where only drinkers and smokers showed an altered susceptibility to disease for this genotype (42). In a study in the Chinese population, *ADH1B* His48Arg was not associated with oesophageal cancer in drinkers or non-drinkers (32). Despite this lack of consistency, one study has shown that the combination of risk SNPs for *ALDH2* Glu504Lys and *ADH1B* His48Arg, together with drinking alcohol and smoking, increases the risk of OSCC synergistically compared with individuals with no risk genotypes who refrain from alcohol consumption and smoking [OR = 189.26 (95% CI = 95.11–376.63)] (8).

In our study, all the significant or suggestive associations detected in cases from the Mixed Ancestry population were also observed in alcohol drinkers but not in non-drinkers. However, the proportion of non-drinkers in this population was very small, with low power to detect any association. Interestingly, the protective effect of the *COX-2* −1195 A > G SNP (P = 0.004) was strengthened in the drinkers subgroup relative to the whole case sample (P = 0.014), again suggestive of a possible gene–environmental interaction, but this requires confirmation by analysis of a larger sample of non-drinking OSCC cases. In relation to smoking status, again all associations observed in the full sample of OSCC cases were also detected in smokers but not in former smokers or non-smokers (those who have never smoked). However, former smokers and non-smokers were a minority in this population, so there was low power to detect associations in these subgroups. The stronger association observed for *MGMT* Leu84Phe with OSCC in smoking cases versus controls (P = 0.003) compared with all cases versus controls (P = 0.023) suggests a possible interaction between smoking and the *MGMT* DNA repair pathway but requires confirmation by analysis of a larger sample of non-smoking cases.

The differences in the results of the genetic association tests between the Black and Mixed Ancestry populations in this study are striking. A number of associations reported previously in European and Asian populations were also detected in the Mixed Ancestry population, albeit with varying levels of support. However, none of the variants tested showed any evidence of association in the Black population. There are several possible explanations for these differences. One is the absence or extreme rarity of the risk SNP in this population, such as was the case for *ADH1B* 48His, *ALDH2* 504Lys and *ADH7* 92Ala. Another is a lack of statistical power to detect modest genetic effects in the samples available. The sample size was larger in the Black population than in the Mixed Ancestry (N = 835 and 628, respectively), and, in several instances, there was high power to detect the same effect seen in the Mixed Ancestry population. However, power in the Black population sample varied substantially, depending on the size of the effect (OR) and the minor allele frequency. Thus, for an SNP with a MAF of 25% and an allelic OR of 1.5, we had 96% power to detect association (at significance level 0.05). However, for *COX-2* −1195A > G (OR 1.34), for example, the MAF was 5% and power was only 30%, whereas for *ALDH2* +82G > A (MAF = 25%, OR = 1.43), the power to detect the same effect seen in the Mixed Ancestry population was >90%. Another possibility is that the SNPs associated with OSCC in other populations may not be the true causal variants and may be merely tagging SNPs, which are in high LD with the causal variant. There is generally a lower level of LD in African populations, and hence, associations may not be detected if the causal variant is not genotyped (43).

A further and potentially important explanation for these results is differences in environmental exposures between these two populations. The significant or suggestive association with OSCC of variants in genes involved in alcohol and acetaldehyde metabolism (*ALDH2* and *ADH1B*) and DNA repair of alkylation-induced mutagenesis (*MGMT*) in the Mixed Ancestry population is consistent with the high rates of smoking and alcohol use in this cohort. However, the proportion of smokers and drinkers was lower in the Black OSCC patients, thus a genetic interaction between, for example, a genetic variant in an aldehyde dehydrogenase gene and alcohol consumption or smoking might be weaker or absent in this population and the genetic association would not be observed. Epidemiological studies in Black South Africans have investigated the contribution of alcohol consumption and smoking to oesophageal cancer risk (44–46), the most recent of which suggests that while alcohol use in itself does

not increase risk (OR 0.9), smoking does (OR 1.9) and the combination of smoking and drinking increases risk further (OR 4.4) (46). However, a major risk factor in this population is residence in the Transkei region of the Eastern Cape Province (47,48), with long-term residence of 35 years or more associated with high risk (OR 14.7) in females (46). Nutritional factors such as deficiency in micronutrients (49,50), and the consumption of maze contaminated with *Fusarium* species, which produce mycotoxins (51) have been proposed as risk factors. Taken together, these studies suggest that dietary factors are likely to be an important component of risk for OSCC in this population. If so, the nature of interacting genetic factors may be substantially different to those found in the Mixed Ancestry and other populations. In view of the success of genome-wide association scans in defining the genetic components of complex disease, this approach is likely to be the most effective means of identifying the relevant genetic pathways involved in the pathogenesis of OSCC in these two populations once sufficient sample sizes for genome-wide studies are available. The identification of such pathways may shed further light on the role of dietary factors in OSCC.

In conclusion, this study provides evidence that several genetic variants in genes involved in alcohol metabolism and DNA repair contribute to genetic susceptibility to OSCC in the Mixed Ancestry but not in the Black population of South Africa. This may be explained both by differences in the genetic history and architecture of these populations and by different environmental exposures.

Supplementary material

Supplementary Tables I and II can be found at http://carcin.oxfordjournals.org/

Funding

Association for International Cancer Research (grant number 09-0625), the Medical Research Council UK, The Generation Trust, the National Institutes of Health Research Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. South African Research Chairs Initiative of the Department of Science and Technology and the National Research Foundation, the International Centre for Genetic Engineering and Biotechnology (ICGEB), the South African Medical Research Council and the University of Cape Town to M.I.P. National Institutes of Health, USA (06565) and the South African Medical Research Council to E.R.

Acknowledgements

We wish to thank Antoinette Olivier and Zenaria Abbas for assisting with the sample collection and processing and the patients and healthy controls for their participation in this study.

Conflict of Interest Statement: None declared.

References

1. Parkin,D.M. *et al*. (2005) Global cancer statistics, 2002. *CA Cancer J. Clin.*, **55**, 74–108.
2. Hendricks,D. *et al*. (2002) Oesophageal cancer in Africa. *IUBMB Life*, **53**, 263–268.
3. Lam,A.K. (2000) Molecular biology of esophageal squamous cell carcinoma. *Crit. Rev. Oncol. Hematol.*, **33**, 71–90.
4. Somdyala,N.I. *et al*. (2010) Cancer incidence in a rural population of South Africa, 1998–2002. *Int. J. Cancer*, **127**, 2420–2429.
5. Lao-Sirieix,P. *et al*. (2010) Genetic predisposition to gastro-oesophageal cancer. *Curr. Opin. Genet. Dev.*, **20**, 210–217.
6. Wang,L.D. *et al*. (2010) Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies susceptibility loci at PLCE1 and C20orf54. *Nat. Genet.*, **42**, 759–763.
7. Abnet,C.C. *et al*. (2010) A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nat. Genet.*, **42**, 764–767.
8. Cui,R. *et al*. (2009) Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. *Gastroenterology*, **137**, 1768–1775.
9. McKay,J.D. *et al*. (2011) A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. *PLoS Genet.*, **7**, e1001333.
10. Li,D. *et al*. (2010) The 341C/T polymorphism in the GSTP1 gene is associated with increased risk of oesophageal cancer. *BMC Genet.*, **11**, 47.
11. Li,D. *et al*. (2005) Association of cytochrome P450 2E1 genetic polymorphisms with squamous cell carcinoma of the oesophagus. *Clin. Chem. Lab. Med.*, **43**, 370–375.
12. Dandara,C. *et al*. (2006) Gene-environment interaction: the role of SULT1A1 and CYP3A5 polymorphisms as risk modifiers for squamous cell carcinoma of the oesophagus. *Carcinogenesis*, **27**, 791–797.
13. de Wit,E. *et al*. (2010) Genome-wide analysis of the structure of the South African coloured population in the Western Cape. *Hum. Genet.*, **128**, 145–153.
14. Li,H. *et al*. (2009) Refined geographic distribution of the oriental ALDH2*504Lys (nee 487Lys) variant. *Ann. Hum. Genet.*, **73**, 335–345.
15. Huang,Q.R. *et al*. (1997) Identification and characterization of polymorphisms in the promoter region of the human Apo-1/Fas (CD95) gene. *Mol. Immunol.*, **34**, 577–582.
16. Wei,S. *et al*. (2010) A single nucleotide polymorphism in the alcohol dehydrogenase 7 gene (alanine to glycine substitution at amino acid 92) is associated with the risk of squamous cell carcinoma of the head and neck. *Cancer*, **116**, 2984–2992.
17. Kimura,Y. *et al*. (2009) A promoter polymorphism in the ALDH2 gene affects its basal and acetaldehyde/ethanol-induced gene expression in human peripheral blood leukocytes and HepG2 cells. *Alcohol Alcohol.*, **44**, 261–266.
18. Yin,M. *et al*. (2010) CASP8 polymorphisms contribute to cancer susceptibility: evidence from a meta-analysis of 23 publications with 55 individual studies. *Carcinogenesis*, **31**, 850–857.
19. Morin,P.A. *et al*. (1999) High-throughput single nucleotide polymorphism genotyping by fluorescent 5' exonuclease assay. *Biotechniques*, **27**, 538–552.
20. Zhang,X. *et al*. (2005) Identification of functional genetic variants in cyclooxygenase-2 and their association with risk of esophageal cancer. *Gastroenterology*, **129**, 565–576.
21. Upadhyay,R. *et al*. (2009) Functional polymorphisms of cyclooxygenase-2 (COX-2) gene and risk for esophageal squmaous cell carcinoma. *Mutat. Res.*, **663**, 52–59.
22. Hall,J. *et al*. (2007) The association of sequence variants in DNA repair and cell cycle genes with cancers of the upper aerodigestive tract. *Carcinogenesis*, **28**, 665–671.
23. Huang,W.Y. *et al*. (2005) Selected genetic polymorphisms in MGMT, XRCC1, XPD, and XRCC3 and risk of head and neck cancer: a pooled analysis. *Cancer Epidemiol. Biomarkers Prev.*, **14**, 1747–1753.
24. Zhong,Y. *et al*. (2010) Effects of O6-methylguanine-DNA methyltransferase (MGMT) polymorphisms on cancer: a meta-analysis. *Mutagenesis*, **25**, 83–95.
25. Sun,T. *et al*. (2007) A six-nucleotide insertion-deletion polymorphism in the CASP8 promoter is associated with susceptibility to multiple cancers. *Nat. Genet.*, **39**, 605–613.
26. Dudbridge,F. (2008) Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum. Hered.*, **66**, 87–98.
27. Crabb,D.W. *et al*. (1989) Genotypes for aldehyde dehydrogenase deficiency and alcohol sensitivity. The inactive ALDH2(2) allele is dominant. *J. Clin. Invest.*, **83**, 314–316.
28. Hashibe,M. *et al*. (2006) Evidence for an important role of alcohol- and aldehyde-metabolizing genes in cancers of the upper aerodigestive tract. *Cancer Epidemiol. Biomarkers Prev.*, **15**, 696–703.
29. Chou,W.Y. *et al*. (1999) An A/G polymorphism in the promoter of mitochondrial aldehyde dehydrogenase (ALDH2): effects of the sequence variant on transcription factor binding and promoter strength. *Alcohol Clin. Exp. Res.*, **23**, 963–968.
30. Stewart,M.J. *et al*. (1998) Binding and activation of the human aldehyde dehydrogenase 2 promoter by hepatocyte nuclear factor 4. *Biochim. Biophys. Acta*, **1399**, 181–186.
31. Ding,J.H. *et al*. (2009) Polymorphisms of alcohol dehydrogenase-2 and aldehyde dehydrogenase-2 and esophageal cancer risk in Southeast Chinese males. *World J. Gastroenterol.*, **15**, 2395–2400.
32. Ding,J.H. *et al*. (2009) Alcohol dehydrogenase-2 and aldehyde dehydrogenase-2 genotypes, alcohol drinking and the risk for esophageal cancer in a Chinese population. *J. Hum. Genet.*, **55**, 97–102.
33. Yokoyama,A. *et al*. (2002) Genetic polymorphisms of alcohol and aldehyde dehydrogenases and glutathione S-transferase M1 and drinking, smoking, and diet in Japanese men with esophageal squamous cell carcinoma. *Carcinogenesis*, **23**, 1851–1859.

34. Sun,T. *et al*. (2004) Polymorphisms of death pathway genes FAS and FASL in esophageal squamous-cell carcinoma. *J. Natl Cancer Inst.*, **96**, 1030–1036.
35. Kase,S. *et al*. (2002) Expression of Fas and Fas ligand in esophageal tissue mucosa and carcinomas. *Int. J. Oncol.*, **20**, 291–297.
36. Gratas,C. *et al*. (1998) Up-regulation of Fas (APO-1/CD95) ligand and down-regulation of Fas expression in human esophageal cancer. *Cancer Res.*, **58**, 2057–2062.
37. Chen,L. *et al*. (2010) CD95 promotes tumour growth. *Nature*, **465**, 492–496.
38. Kanemitsu,S. *et al*. (2002) A functional polymorphism in fas (CD95/APO-1) gene promoter associated with systemic lupus erythematosus. *J. Rheumatol.*, **29**, 1183–1188.
39. Sibley,K. *et al*. (2003) Functional FAS promoter polymorphisms are associated with increased risk of acute myeloid leukemia. *Cancer Res.*, **63**, 4327–4330.
40. Haiman,C.A. *et al*. (2008) A promoter polymorphism in the CASP8 gene is not associated with cancer risk. *Nat. Genet.*, **40**, 259–260; author Reply, 260–261.
41. Lewis,S.J. *et al*. (2005) Alcohol, ALDH2, and esophageal cancer: a meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach. *Cancer Epidemiol. Biomarkers Prev.*, **14**, 1967–1971.
42. Hashibe,M. *et al*. (2008) Multiple ADH genes are associated with upper aerodigestive cancers. *Nat. Genet.*, **40**, 707–709.
43. Teo,Y.Y. *et al*. (2010) Methodological challenges of genome-wide association analysis in Africa. *Nat. Rev. Genet.*, **11**, 149–160.
44. McGlashan,N.D. *et al*. (1982) Cancer of the oesophagus and the use of tobacco and alcoholic beverages in Transkei, 1975-6. *Int J Cancer*, **29**, 249–256.
45. Segal,I. *et al*. (1988) Factors associated with oesophageal cancer in Soweto, South Africa. *Br. J. Cancer*, **58**, 681–686.
46. Pacella-Norman,R. *et al*. (2002) Risk factors for oesophageal, lung, oral and laryngeal cancers in black South Africans. *Br. J. Cancer*, **86**, 1751–1756.
47. Rose,E.F. (1973) Esophageal cancer in the Transkei: 1955-69. *J. Natl Cancer Inst.*, **51**, 7–16.
48. Burrell,R.J. (1962) Esophageal cancer among Bantu in the Transkei. *J. Natl Cancer Inst.*, **28**, 495–514.
49. Van Rensburg,S.J. (1981) Epidemiologic and dietary evidence for a specific nutritional predisposition to esophageal cancer. *J. Natl Cancer Inst.*, **67**, 243–251.
50. Van Rensburg,S.J. *et al*. (1983) Nutritional status of African populations predisposed to esophageal cancer. *Nutr. Cancer*, **4**, 206–216.
51. Shephard,G.S. *et al*. (2007) Exposure assessment for fumonisins in the former Transkei region of South Africa. *Food Addit. Contam.*, **24**, 621–629.

Received July 27, 2011; revised August 25, 2011; accepted September 10, 2011

Supplementary Table I: HGVS nomenclature for the OSCC variants tested in this study

Gene	dbSNP ID	Common variant name*	Updated name (HGVS/NCBI)	Location
ALDH2	rs671	Glu504Lys (G>A)	p.504E>K (NP_000681.2)	coding
ALDH2	rs886205	+82 A>G	c.-360A>G (NM_000690.2)	5'UTR
ALDH2	rs441	-261 C>T	c.682-261C>T (NM_000690.2)	intron 6
ADH1B	rs1229984	Arg48His (G>A)	p.48R>H (NP_000659.2)	coding
ADH7	rs1573496	Gly92Ala	p.92G>A (NP_000664.2)	coding
FAS	rs1800682	-670 A>G	c.-671A>G (NM_000043.3)	upstream
FAS	rs2234767	-1377 G>A	c.-1378G>A (NM_000043.3)	upstream
FASL	rs763110	-844 T>C	c.-844T>C (NM_000639.1)	upstream
CASP8	rs1045485	Asp302His (G>C)	p.302D>H (NP_001219.2)	coding
CASP8	rs3834129	-652 6N ins/del	g.4352_4357del6 (NG_007497.1)	upstream
COX-2	rs20417	-765 G>C	c.-899G>C (NM_000963.2)	upstream
COX-2	rs689466	-1195 G>A	c.-1329G>A (NM_000963.2)	upstream
MGMT	rs12917	Leu84Phe	p.115L>F (NP_002403.2)	coding

Supplementary Table II: Association of polymorphisms with OSCC in the South African Black and Mixed Ancestry populations

	Black population				Mixed Ancestry population			
	Cases (%)	Controls (%)	OR (95% CI)	P-value	Cases (%)	Controls (%)	OR (95% CI)	P-value
ADH1B, Arg48His								
G/G	-	-			176 (89.8)	344 (81.1)		Ref
G/A	-	-	-	-	19 (9.7)	77 (18.2)	0.48 (0.28 - 0.82)	0.007
A/A	-	-	-	-	1 (0.5)	3 (0.7)	0.65 (0.07 - 6.31)	0.709
G					371 (94.6)	765 (90.2)		Ref
A	-	-	-	-	21 (5.4)	83 (9.8)	0.52 (0.32 - 0.86)	0.009
ADH7, Gly92Ala								
C/C	272 (100.0)	414 (99.8)			143 (97.3)	404 (96.0)		Ref
C/G	0 (0.0)	1 (0.2)			4 (2.7)	17 (4.0)	0.66 (0.22 - 2.01)	0.466
G/G	0 (0.0)	0 (0.0)			0 (0.0)	0 (0.0)		
C	544	829 (99.9)			290 (98.6)	825 (98.0)		Ref
G	0	1 (0.1)			4 (1.4)	17 (2.0)	0.70 (0.22 - 2.01)	0.471
ALDH2,+82 G>A								
G/G	191 (55.8)	258 (56.0)	Ref	Ref	73 (36.9)	120 (28.2)	Ref	Ref
G/A	133 (38.9)	174 (37.7)	1.03 (0.77 - 1.38)	0.831	91 (46.0)	194 (45.6)	0.77 (0.53 - 1.13)	0.183
A/A	18 (5.3)	29 (6.3)	0.84 (0.45 - 1.55)	0.575	34 (17.2)	111 (26.1)	0.50 (0.31 - 0.82)	0.005
G	515 (75.3)	690 (74.8)	Ref	Ref	237 (59.8)	434 (51.1)	Ref	Ref
A	169 (24.7)	232 (25.2)	0.98 (0.78 - 1.23)	0.835	159 (40.2)	416 (48.9)	0.70 (0.55 - 0.89)	0.004
ALDH2, -261 T>C								
T/T	243 (71.3)	340 (73.0)	Ref	Ref	128 (66.0)	275 (66.1)	Ref	Ref
T/C	91 (26.7)	117 (25.1)	1.09 (0.79 - 1.50)	0.604	62 (32.0)	121 (29.1)	1.10 (0.76 - 1.60)	0.612
C/C	7 (2.1)	9 (1.9)	1.09 (0.40 - 2.96)	0.868	4 (2.1)	20 (4.8)	0.43 (0.14 - 1.28)	0.120
T	577 (84.6)	797 (85.5)	Ref	Ref	318 (82.0)	671 (80.6)	Ref	Ref
C	105 (15.4)	135 (14.5)	1.07 (0.81 - 1.42)	0.611	70 (18.0)	161 (19.4)	0.92 (0.67 - 1.25)	0.587
COX-2, -765 G>C								
G/G	100 (28.8)	110 (23.8)	Ref		81 (42.6)	195 (46.2)	Ref	Ref
G/C	167 (48.1)	230 (49.8)	0.80 (0.57 - 1.12)	0.190	75 (39.5)	183 (43.4)	0.99 (0.68 - 1.43)	1.000
C/C	80 (23.1)	122 (26.4)	0.72 (0.49 - 1.07)	0.101	34 (17.9)	44 (10.4)	1.86 (1.11 - 3.12)	0.018
G	367 (52.9)	450 (48.7)	Ref	Ref	237 (62.4)	573 (67.9)	Ref	Ref
C	327 (47.1)	474 (51.3)	0.85 (0.69 - 1.03)	0.096	143 (37.6)	271 (32.1)	1.28 (0.99 - 1.64)	0.059
COX-2,-1195 A>G								
A/A	301 (87.2)	417 (89.7)			154 (79.4)	298 (71.1)	Ref	Ref
A/G	44 (12.8)	47 (10.1)	1.30 (0.84 - 2.01)	0.243	40 (20.6)	112 (26.7)	0.69 (0.46 - 1.04)	0.077
G/G	0 (0.0)	1 (0.2)	-		0 (0.0)	9 (2.1)	-	
A	646 (93.6)	881 (94.7)	Ref	Ref	348 (89.7)	708 (84.5)	Ref	Ref
G	44 (6.4)	49 (5.3)	1.22 (0.80 - 1.86)	0.343	40 (10.3)	130 (15.5)	0.63 (0.43 - 0.91)	0.014
MGMT, Leu84Phe								
C/C	225 (65.0)	300 (64.0)	Ref	Ref	120 (61.2)	294 (69.5)	Ref	Ref
C/T	111 (32.1)	155 (33.0)	0.95 (0.71 - 1.29)	0.762	65 (33.2)	116 (27.4)	1.37 (0.95 - 1.99)	0.093
T/T	10 (2.9)	14 (3.0)	0.95 (0.42 - 2.18)	1.000	11 (5.6)	13 (3.1)	2.07 (0.90 - 4.76)	0.080
C	561 (81.1)	755 (80.5)	Ref	Ref	305 (77.8)	704 (83.2)	Ref	Ref
T	131 (18.9)	183 (19.5)	0.96 (0.75 - 1.24)	0.770	87 (22.2)	142 (16.8)	1.41 (1.05 - 1.90)	0.023
CASP8,Asp302His								
G/G	247 (71.8)	332 (71.6)	Ref	Ref	135 (70.3)	326 (77.3)	Ref	Ref
G/C	88 (25.6)	123 (26.5)	0.96 (0.70 - 1.32)	0.810	49 (25.5)	86 (20.4)	1.38 (0.92 - 2.06)	0.121
C/C	9 (2.6)	9 (1.9)	1.34 (0.53 - 3.44)	0.535	8 (4.2)	10 (2.4)	1.93 (0.75 - 5.00)	0.168
G	582 (84.6)	787 (84.8)	Ref	Ref	319 (83.1)	738 (87.4)	Ref	Ref
C	106 (15.4)	141 (15.2)	1.02 (0.77 - 1.34)	1.000	65 (16.9)	106 (12.6)	1.42 (1.01 - 1.98)	0.040
CASP8, -652 6N ins/del								
Ins / Ins	82 (24.9)	112 (24.5)	Ref	Ref	76 (39.8)	160 (38.6)	Ref	Ref
Del / Ins	153 (46.5)	232 (50.7)	0.90 (0.63 - 1.27)	0.559	83 (43.5)	188 (45.4)	0.92 (0.63 - 1.35)	0.703
Del / Del	94 (28.6)	114 (24.9)	1.12 (0.75 - 1.67)	0.555	32 (16.8)	66 (15.9)	1.02 (0.61 - 1.68)	1.000
Ins	317 (48.2)	456 (49.8)	Ref	Ref	235 (61.5)	508 (61.4)	Ref	Ref
Del	341 (51.8)	460 (50.2)	1.06 (0.87 - 1.30)	0.530	147 (38.5)	320 (38.6)	0.99 (0.77 - 1.27)	1.000
FAS, -670 G>A								
G/G	210 (61.2)	288 (61.8)	Ref	Ref	87 (44.6)	155 (36.9)	Ref	Ref
G/A	116 (33.8)	146 (31.3)	1.09 (0.81 - 1.47)	0.577	77 (39.5)	189 (45.0)	0.73 (0.50 - 1.05)	0.092
A/A	17 (5.0)	32 (6.9)	0.73 (0.39 - 1.35)	0.311	31 (15.9)	76 (18.1)	0.73 (0.44 - 1.19)	0.204
G	536 (78.1)	722 (77.5)	Ref	Ref	251 (64.4)	499 (59.4)	Ref	Ref
A	150 (21.9)	210 (22.5)	0.96 (0.76 - 1.22)	0.750	139 (35.6)	341 (40.6)	0.81 (0.63 - 1.04)	0.097
FAS, -1377G>A								
G/G	276 (81.4)	393 (86.2)	Ref	Ref	144 (74.2)	278 (67.3)	Ref	Ref
G/A	61 (18.0)	60 (13.2)	1.45 (0.98 - 2.13)	0.061	46 (23.7)	119 (28.8)	0.75 (0.50 - 1.11)	0.146
A/A	2 (0.6)	3 (0.7)	0.95 (0.16 - 5.72)	1.000	4 (2.1)	16 (3.9)	0.48 (0.16 - 1.47)	0.191
G	613 (90.4)	846 (92.8)	Ref	Ref	334 (86.1)	675 (81.7)	Ref	
A	65 (9.6)	66 (7.2)	1.36 (0.95 - 1.94)	0.092	54 (13.9)	151 (18.3)	0.72 (0.52 - 1.01)	0.058
FASL, -844 T>C								
T/T	216 (64.7)	297 (64.6)	Ref	Ref	63 (33.0)	159 (38.0)	Ref	Ref
C/T	108 (32.3)	152 (33.0)	0.98 (0.72 - 1.32)	0.880	97 (50.8)	195 (46.7)	1.26 (0.86 - 1.84)	0.240
C/C	10 (3.0)	11 (2.4)	1.25 (0.52 - 3.00)	0.616	31 (16.2)	64 (15.3)	1.22 (0.73 - 2.05)	0.448
T	540 (80.8)	746 (81.1)	Ref	Ref	223 (58.4)	513 (61.4)	Ref	Ref
C	128 (19.2)	174 (18.9)	1.02 (0.79 - 1.31)	1.000	159 (41.6)	323 (38.6)	1.13 (0.89 - 1.45)	0.323

3.3.1 Additional analysis: Gene-environment interactions

Since publication of the paper, demographic information for the South African Black and Mixed ancestry controls became available, as shown in Table 3.1.

Table 3.1: Demographic information for cases and controls

	Black population		Mixed Ancestry population	
	Cases	**Controls**	**Cases**	**Controls**
Total	358	477	201	427
Age, mean years (SD)	59.8 (11.3)	56.9 (14.7)	60.5 (10.6)	57.7 (14.4)
Sex, *n* (%):				
Male	182 (50.8)	183 (38.4)	131 (65.2)	122 (28.6)
Female	176 (49.2)	293 (61.4)	70 (34.8)	300 (70.3)
Unknown	0	1 (0.2)	0	5 (1.2)
Smoking status, *n* (%):				
Smoker	228 (63.7)	177 (37.1)	189 (94.0)	269 (63.0)
Non-smoker	130 (36.3)	294 (61.6)	10 (5.0)	158 (37.0)
Unknown	0	6 (1.3)	2 (1.0)	0
Alcohol consumption, *n* (%):				
Drinker	228 (63.7)	262 (54.9)	164 (81.6)	184 (43.1)
Non-drinker	128 (35.8)	214 (44.9)	37 (18.4)	242 (56.7)
Unknown	2 (0.6)	1 (0.2)	0	1 (0.2)

Therefore, we have used this information to carry out further analysis of possible gene-environment interactions for those variants with a significant or suggested association with OSCC ($P<0.05$). Due to the low number of non-smokers in the Mixed Ancestry population, only alcohol analysis was carried out, see Table 3.2 for results.

Table 3.2: Gene-alcohol interaction tests
Tests are for the variants which showed evidence of association (P<0.05) with OSCC in the South African Mixed Ancestry population. Includes stratified analysis of alcohol consumption, a case-only analysis, and a gene-environment (G x E) interaction test using logistic regression controlling for age, sex, tobacco use and alcohol consumption.

Variant	Minor allele	Case-control: Drinkers only				Case-control: Non-drinkers only			
		MAF: Cases	MAF: Controls	OR (95% CI)	P-value	MAF: Cases	MAF: Controls	OR (95% CI)	P-value
ADH1B Arg48His	T	0.056	0.086	0.63 (0.35 - 1.15)	0.1328	0.042	0.104	0.38 (0.11 - 1.24)	0.0949
COX-2 -1195 A>G	G	0.089	0.157	0.52 (0.32 - 0.85)	0.0079	0.162	0.153	1.07 (0.55 - 2.09)	0.8341
CASP8 Asp302His	C	0.183	0.108	1.84 (1.19 - 2.85)	0.0060	0.111	0.138	0.78 (0.36 - 1.71)	0.5400
ALDH2 +82 G>A	A	0.394	0.503	0.64 (0.48 - 0.87)	0.0045	0.432	0.481	0.82 (0.50 - 1.34)	0.4320
MGMT Leu84Phe	T	0.233	0.175	1.43 (0.98 - 2.08)	0.0619	0.176	0.162	1.10 (0.58 - 2.11)	0.7643

Variant	Minor allele	Case-only				G x E	
		MAF: Drinkers	MAF: Non-drinkers	OR (95% CI)	P-value	OR (95% CI)	P-value
ADH1B Arg48His	T	0.056	0.042	1.37 (0.39 - 4.78)	0.6195	2.17 (0.54 - 8.72)	0.2733
COX-2 -1195 A>G	G	0.089	0.162	0.51 (0.24 - 1.05)	0.0632	0.51 (0.21 - 1.24)	0.1367
CASP8 Asp302His	C	0.183	0.111	1.79 (0.81 - 3.94)	0.1443	2.39 (0.96 - 5.94)	0.0608
ALDH2 +82 G>A	A	0.394	0.432	0.85 (0.51 - 1.43)	0.5474	0.98 (0.54 - 1.79)	0.9502
MGMT Leu84Phe	T	0.233	0.176	1.42 (0.74 - 2.73)	0.2876	1.43 (0.66 - 3.09)	0.3675

The case-control interaction analysis suggested that several variants are associated with OSCC in drinkers but not in non-drinkers. The association in drinkers became more significant than the overall association for two variants: *COX-2* -1195 A>G (drinkers P=0.0079; overall P=0.014) and *CASP8* Asp302His (drinkers P=0.006; overall P=0.040). However, the case-only analysis does not identify significant differences between drinkers and non-drinkers for any of the variants, and nor does the interaction test using logistic regression. This is perhaps due to the low number of non-drinkers in this population (n= 37, 18.4% of samples), and would probably require a much larger sample size to determine whether any gene-environmental interactions exist.

In summary, the revised interaction analysis using smoking and drinking data from controls is consistent with the analysis reported in our paper.

3.4 Summary

This chapter investigated genetic susceptibility to OSCC in the South African Black and Mixed Ancestry populations, focusing on 13 variants in 8 genes with robust evidence of association in other populations. A SNP in the promoter region of *ALDH2* (+82 G>A, rs886205) was associated with a reduced risk of OSCC in the Mixed Ancestry population. None of the variants were associated with the disease in the Black population. This study suggests differences in the genetic contribution to OSCC both between the Black and Mixed Ancestry populations, and between the South Africans and other high risk populations.

4 Distinct genetic association at the *PLCE1* locus with oesophageal squamous cell carcinoma in the South African population

The work in this chapter was published in *Carcinogenesis* in 2012 (Bye *et al.* 2012). The paper is included as published on pages 132-146. A brief introduction to the area of work is given below.

4.1 Genome-wide association studies for OSCC in the Chinese population

Three independent OSCC genome-wide association studies were published during 2010 and 2011 (Abnet *et al.* 2010; Wang *et al.* 2010.a; Wu *et al.* 2011.c). All studies consisted of over 1,000 cases and 1,000 controls in the original GWAS, with replication phases ranging from ~2,100 to ~7,600 cases and ~3,300 to ~11,000 controls. In total, eight variants in six loci were associated with OSCC at genome-wide significance ($P<5 \times 10^{-8}$). Of these, five were novel loci, with the known loci located on chromosome 12q24, a region containing *ALDH2*. This region has been investigated in many candidate gene association studies including in the South African Black and Mixed Ancestry populations, see Chapter 3. The novel susceptibility loci are summarized in Table 4.1 and are discussed below.

Table 4.1: Summary of OSCC genome-wide association results in Chinese populations

Gene/locus	Variant	Location	Author	Alleles (major, minor)	MAF: Cases	MAF: Controls	OR (95% CI)	P-value
PLCE1	rs2274223	Exonic	Wang *et al.* 2010.a	A, G	0.27	0.2	1.43 (1.37 - 1.49)	7.46×10^{-56}
			Abnet *et al.* 2010		0.259	0.209	1.34 (1.22 - 1.48)	3.85×10^{-9}
			Wu *et al.* 2011.c		0.26	0.21	1.34 (1.26 - 1.42)	3.73×10^{-20}
C20orf54/*SLC52A3*	rs13042395	Upstream	Wang *et al.* 2010.a	C, T	0.24	0.32	0.66 (0.58 - 0.74)	1.21×10^{-11}
near *UNC5CL*	rs10484761	Downstream	Wu *et al.* 2011.c	A, G	0.12	0.09	1.33 (1.23 - 1.45)	7.48×10^{-12}
PDE4D	rs10052657	Intronic	Wu *et al.* 2011.c	C, A	0.08	0.12	0.67 (0.62 - 0.73)	1.97×10^{-19}
RUNX1	rs2014300	Intronic	Wu *et al.* 2011.c	G, A	0.12	0.17	0.70 (0.65 - 0.75)	8.09×10^{-22}

PLCE1 rs2274223 was the only variant to be associated in more than one GWAS, being identified in all three studies, thus providing strong evidence that it is a true susceptibility locus in this population. This was also the only associated variant to be located in an exon, producing an amino acid change from histidine to arginine at position 1927. *PLCE1* encodes phospholipase C epsilon 1, which converts phosphatidyl-inositol 4,5-biphosphate to form diacylglycerol and inositol 1,4,5-triphosphate, which leads to the activation of protein kinase C. Further insight into the functions of the protein is explored in the discussion section of the published paper. Before the identification of *PLCE1* variants as susceptibility loci for OSCC, *PLCE1* mRNA expression was known to be reduced in certain tumour tissue compared to normal tissue, for example, in colorectal cancer (Sorli *et al.* 2005; Wang *et al.* 2008; Danielsen *et al.* 2011; Wang *et al.* 2012). In addition, the corresponding PLCE1 protein level was also found to be decreased in tumour tissue (Wang *et al.* 2012). Together with the observation that overexpression of *PLCE1* inhibited proliferation of colon cancer cells, *PLCE1* was suggested to have a suppressive role in the development of this cancer (Wang *et al.* 2012). In OSCC, *PLCE1* mRNA was also reduced in tumour tissue compared to normal tissue (Hu *et al.* 2012), but levels of protein production were conflicting with one study showing increased levels (Wang *et al.* 2010.a) and another having no significant changes compared to normal tissue (Hu *et al.* 2012). This is in contrast to the colorectal cancer study (Wang *et al.* 2012).

Several studies have attempted to replicate the association of *PLCE1* rs2274223 with OSCC. In two additional Chinese populations, one from a high-risk disease region and one from a lower-risk region, the variant was associated with an increased risk of OSCC in a recessive model for GG vs. AA genotypes (OR = 1.95, 95% CI = 1.05 – 3.59, P=0.034; and OR = 1.49, 95% CI = 1.03-2.17; P=0.037, for each population respectively) (Gu *et al.* 2012; Hu *et al.* 2012). Neither of these studies report the allelic association result which is the analysis model used in GWAS. Calculating this from the number of each genotype stated

in the papers leads to allelic associations of P=0.0055 (OR =1.40, 95% CI = 1.10-1.78) and P=0.0018 (OR=1.25; 95% CI =1.08-1.43) for the high and low-risk regions of China, respectively. In a Caucasian USA study of 52 OSCC cases and 211 controls, the rs2274223 GG genotype was reported as being protective against the disease in a dominant model (OR = 0.5, 95% CI = 0.3 – 1.0 for GG vs. AA genotype), which is the opposite effect observed in the Chinese populations (Palmer *et al.* 2012). However, the authors did not report p-values and they did not account for multiple testing. Calculating the allelic association using genotype numbers reported, no significant association with disease was identified (P=0.352; OR = 0.82, 95% CI = 0.53-1.25). Whether this variant has an opposite effect on OSCC susceptibility in Caucasians compared to Chinese populations needs to be further explored using a much larger sample size since the published study would have very low power to detect the effect sizes reported in the Chinese studies.

RUNX1 (*Runt-related transcription factor 1*) encodes for a transcription factor which, when part of a protein complex, is able to regulate the transcription of proteins involved in growth, survival and differentiation pathways (reviewed in Blyth *et al.* 2005). The protein is known to be essential for haematopoiesis (Okuda *et al.* 1996), and is frequently reported to be involved in the development of a variety of leukaemias through chromosomal translocations and point mutations. For example, the *RUNX1-ETO* chromosome rearrangement is present in 10-20% of acute myeloid leukaemias (Blyth *et al.* 2005). The fusion proteins produced from chromosomal translocations are thought to act as dominant negative inhibitors of expression of the normal *RUNX1* gene and may also acquire novel functions (Blyth *et al.* 2005). *RUNX1* is also downregulated in solid cancers, including gastric cancer (Sakakura *et al.* 2005). In addition to this tumour suppressive role, *RUNX1*, like other *RUNX* genes, is proposed to also act as an oncogene in a context-dependent manner (reviewed in Blyth *et al.* 2005). This is supported by the amplification or increased expression of *RUNX1* in several cancers, including childhood acute

lymphoblastic leukemia (reviewed in Roumier *et al.* 2003) and endometrial carcinoma (Planaguma *et al.* 2004). In a small study of OSCC patients, *RUNX1* expression was found not to be statistically different between tumour and normal tissue (Tonomoto *et al.* 2007).

This complex role of *RUNX1* in cancer development is probably due to its ability to both activate and repress transcription of genes involved in key pathways of growth and differentiation, and hence the function of downstream targets needs to be explored (reviewed in Blyth *et al.* 2005). In addition to regulating genes in the haematopoiesis pathway, *RUNX1* may also be involved in the TGFβ and p53 signalling pathways in myeloid cells (reviewed in Blyth *et al.* 2005). If this is also the case in other cell types, this may provide a potential link between *RUNX1* and solid cancer. More recently, *RUNX1* has also been implicated in immune response pathways, such as regulating interleukin 17-producing helper T cell differentiation (Lazarevic *et al.* 2011), which may also be relevant to tumour formation.

The *RUNX1* rs2014300 variant associated with OSCC is located within an intron and has an unknown function. No studies have attempted to replicate the original GWAS finding in OSCC or other cancers.

Another variant associated with OSCC in one of these GWAS was rs13042395 which is located downstream of *C20orf54*, also known as *SLC52A3*, a riboflavin transporter (Wang *et al.* 2010.a). Riboflavin, or vitamin B2, is essential in cellular homeostasis for normal growth and development, and also plays a role in carbohydrate, lipid and amino acid metabolism (Subramanian *et al.* 2011). Riboflavin deficiency has been associated with an increased risk of OSCC in some studies (Siassi and Ghadirian 2005) but not others (Siassi *et al.* 2000), with riboflavin supplementation possibly leading to a reduced incidence of OSCC (Blot *et al.* 1993; He *et al.* 2009). The function of the associated rs13042395 variant is unknown. Candidate gene association studies have

attempted to replicate the GWAS finding without success: the variant was not associated with OSCC in another Chinese population or a Caucasian population from USA (Gu *et al.* 2012; Palmer *et al.* 2012). In the latter study particularly, this may be due to an under-powered study as sample sizes were very small with only 52 OSCC cases and 211 controls (Palmer *et al.* 2012). Other variants in *C20orf54* have also been tested for association with OSCC, with Ji *et al*. (2011) identifying two functional variants in this gene associated with the disease.

PDE4D is a c-AMP phosphodiesterase, responsible for the hydrolysis of cAMP to form 5'AMP. cAMP is involved in several cellular signalling pathways including proliferation and apoptosis (Savai *et al.* 2010). *PDE4D* is proposed to be a tumour suppressor gene based on deletions observed in oesophageal adenocarcinoma and lung cancer (Weir *et al.* 2007; Nancarrow *et al.* 2008; Gu *et al.* 2010). However, overexpression of *PDE4D* in lung tumours resulted in increased cell proliferation, with decreased proliferation observed when PDE4D was inhibited (Pullamsetti *et al.* 2012). These conflicting reports may be due to cAMP having both pro-apoptotic and anti-apoptotic roles (reviewed in Insel *et al.* 2012). The *PDE4D* rs10052657 variant associated with OSCC in the study by Wu *et al.* (2011) has not been tested for association in any other populations or with different cancer subtypes. It is located within an intron, with an unknown functional effect.

Finally, a variant on chromosome 6p21 was found to be associated with OSCC in a region which contains no known genes. The nearest gene, which is located about 200 Kb away, is *UNC5CL,* which inhibits activation of the NF-kB transcription factor and hence, may regulate downstream pathways of NF-kB including cell proliferation. The variant was not associated with head and neck cancer in a Chinese population (Yuan *et al.* 2013), and has not been further investigated in OSCC.

4.2 Association studies in South African populations

The aim of this chapter was to investigate the novel loci associated with OSCC in the Chinese GWAS studies with genetic susceptibility to OSCC in the South African Black and Mixed Ancestry populations. Although not all of the loci currently have a known role which would indicate an involvement in cancer development, future functional work may establish this. This work is published in my paper entitled “Distinct genetic association at the *PLCE1* locus with oesophageal squamous cell carcinoma in the South African population”, which is presented in journal format overleaf. Details of the Materials and Methods are given in Chapter 2.

My contribution to this paper has been to perform all the lab work, including the genotyping and Sanger sequencing, and the analysis of results, as well as writing the paper, together with C. Mathew.

Carcinogenesis vol.33 no.11 pp.2155–2161, 2012
doi:10.1093/carcin/bgs262
Advance Access publication August 6, 2012

Distinct genetic association at the PLCE1 locus with oesophageal squamous cell carcinoma in the South African population

Hannah Bye[1], Natalie J.Prescott[1], Cathryn M.Lewis[1], Marco Matejcic[2], Loven Moodley[3], Barbara Robertson[4], Christo van Rensburg[5], M.Iqbal Parker[2] and Christopher G.Mathew[1,*]

[1]Department of Medical and Molecular Genetics, King's College London, King's Health Partners, Guy's Hospital, London, United Kingdom, [2]International Centre for Genetic Engineering and Biotechnology and Division of Medical Biochemistry, University of Cape Town, Cape Town, South Africa, [3]Chris Barnard Division of Cardiothoracic Surgery, University of Cape Town, Cape Town, South Africa, [4]Department of Radiation Oncology, University of Cape Town, Cape Town, South Africa, [5]Division of Gastroenterology and Hepatology, Tygerberg Academic Hospital and Stellenbosch University, Tygerberg, South Africa

*To whom correspondence should be addressed. Tel: +44 (0) 20 7188 3713; Fax: +44 (0) 207 188 2585
Email: christopher.mathew@kcl.ac.uk

Oesophageal squamous cell carcinoma (OSCC) has a high prevalence in the Black and Mixed Ancestry populations of South Africa. Recently, three genome-wide association studies in Chinese populations identified five new OSCC susceptibility loci, including variants at *PLCE1*, *C20orf54*, *PDE4D*, *RUNX1* and *UNC5CL*, but their contribution to disease risk in other populations is unknown. In this study, we report testing variants from these five loci for association with OSCC in the South African Black (407 cases and 849 controls) and Mixed Ancestry (257 cases and 860 controls) populations. The *RUNX1* variant rs2014300, which reduced risk in the Chinese population, was associated with an increased risk of OSCC in the Mixed Ancestry population [odds ratio (OR) = 1.33, 95% confidence interval (CI) = 1.09–1.63, *P* = 0.0055], and none of the five loci were associated in the Black population. Since *PLCE1* variants increased the risk of OSCC in all three Chinese studies, this gene was investigated further by sequencing in 46 Black South Africans. This revealed 48 variants, 10 of which resulted in amino acid substitutions, and much lower linkage disequilibrium across the *PLCE1* locus than in the Chinese population. We genotyped five *PLCE1* variants in cases and controls, and found association of Arg548Leu (rs17417407) with a reduced risk of OSCC (OR = 0.74, 95% CI = 0.60–0.93, *P* = 0.008) in the Black population. These findings indicate several differences in the genetic contribution to OSCC between the South African and Chinese populations that may be related to differences in their genetic architecture.

Introduction

Oesophageal cancer is the eighth most common cancer worldwide and the sixth most common form of death from cancer (1). The predominant subtype in developing countries is oesophageal squamous cell carcinoma (OSCC), whereas oesophageal adenocarcinoma is more common in the western world where its incidence is increasing. High-risk regions for OSCC include southern Africa, Japan, China and northern Iran. In southern Africa, oesophageal cancer is the third most common cancer in both males and females with age-adjusted incidence rates of 22.3 and 11.7 per 100 000, respectively (2); higher rates are observed in certain regions such as the Eastern Cape Province of South Africa (3). Environmental risk factors for the development of OSCC in South Africa include alcohol intake and tobacco use, nutritional deficiencies, consumption of *Fusarium*-(fungi) contaminated maize and infection with human papilloma virus (reviewed in ref. 4).

Candidate gene studies have analysed multiple genetic variants for association with OSCC in two indigenous South African populations, the Mixed Ancestry and the Black populations, with a high prevalence of this disease. Single nucleotide polymorphisms (SNPs) in *GSTM1*, *GSTP1* (5), *CYP3A5* (6) and *CYP2E1* (7) showed some evidence of association, although sample sizes were small and data from the two populations were pooled in some analyses. In view of the differences in population structure between these two South African populations (8), recent studies by our group have analysed the Black and Mixed Ancestry populations separately, and increased the power to detect association by expansion of the sample sizes. We reported significant association of a 37 kb deletion in *GSTT2B* (9) and of the variant *ALDH2* + 82 G>A (rs886205) (10) with OSCC in the Mixed Ancestry population. However, none of the variants tested in these studies were associated with OSCC in the Black South African population, which may be related to differences in their ancestry or environmental exposures (10), or to chance.

The development of genome-wide association scans (GWAS) has had a major impact on the discovery of susceptibility genes for complex disease. The first GWAS for OSCC was published in 2009 and identified significant associations with *ALDH2* Glu504Lys (rs671) on chromosome 12q24 and *ADH1B* Arg48His (rs1229984) on chromosome 4q23 in the Japanese population (11). These SNPs were tested for association in the South African Black and Mixed Ancestry populations; the *ALDH2* Glu504Lys variant was non-polymorphic in both populations, and *ADH1B* Arg48His showed a suggestive association in the Mixed Ancestry population but was non-polymorphic in the Black population (10). Recently, three independent OSCC GWAS in Chinese populations have identified a total of eight SNPs in six susceptibility loci, including *PLCE1* His1927Arg (rs2274223) on chromosome 10q23, which was the only locus significantly associated in all three studies (12–14). Other loci identified were *C20orf54/SLC52A3* (rs13042395) on chromosome 20p13 (13), *PDE4D* (rs10052657) on chromosome 5q12 (14), *RUNX1* (rs2014300) on chromosome 21q22.3 (14), a variant near *UNC5CL* (rs10484761) on chromosome 6p21.1 (14), and three SNPs at a locus on 12q24—*ACAD10* (rs11066015), *C12orf51* (rs2074356) and rs11066280 (14). The contribution of the new loci to the risk of OSCC in other populations is unknown.

The aim of this study was to determine whether the new loci identified in the Chinese GWAS were associated with OSCC in the South African Black and Mixed Ancestry populations, and to investigate genetic variation and the genetic architecture of the *PLCE1* gene in the South African population.

Materials and methods

Study subjects

This study consisted of 407 OSCC patients and 849 controls from the South African Black population, and 257 OSCC patients and 860 controls from the South African Mixed Ancestry population. The Black patients were mainly Xhosa-speakers (98.8%) from the Eastern or Western Cape of South Africa. The Black controls were recruited from factories and outpatient clinics in the Western Cape. The proportion of Xhosa-speakers was 98.2%; they had no history of any cancer, lived in the same residential areas, and had a similar socio-economic status to the patients. The Mixed Ancestry cases and controls were all recruited from the Western Cape. This is an admixed population with major ancestral components from the indigenous Khoisan, Bantu-speaking Africans, Europeans and Asians (8). Patients with histologically confirmed primary invasive OSCC were recruited between March 2000 and August 2011 at Groote Schuur and Tygerberg Hospitals in Cape Town. Data on alcohol consumption and tobacco use were available for both cases and controls. Smoking status

Abbreviations: CI, confidence interval; GWAS, genome-wide association scans; LD, linkage disequilibrium; MAF, minor allele frequency; OR, odds ratio; OSCC, oesophageal squamous cell carcinoma; SNP, single nucleotide polymorphisms; UTR, untranslated region.

was subdivided into ever-smokers (those who had smoked at some point in their lives) or never-smokers. Drinkers were defined as subjects who consumed alcohol at least once every week. Demographic and exposure data are given in Table I. Whole blood samples were collected with informed consent from all subjects and DNA was extracted at the University of Cape Town. Ethical approval for the study was obtained from the joint University of Cape Town/Groote Schuur Hospital Research Ethics Committee and the University of Stellenbosch/Tygerberg Hospital Ethics Committee.

SNP selection and genotyping

Index SNPs (those with the strongest evidence for association with OSCC in the Chinese GWAS studies) from five of the six loci were polymorphic in the Yoruban (Nigerian) and Masai in Kinyawa (Kenyan) HapMap populations, suggesting that they were likely to be informative in the Black South African population. These five SNPs, *PLCE1* His1927Arg (rs2274223), *C20orf54/SLC52A3* (rs13042395), *PDE4D* (rs10052657), *RUNX1* (rs2014300) and a variant near *UNC5CL* (rs10484761), were selected for genotyping. Two of the SNPs at the 12q24 locus (rs2074356 and rs11066280) were non-polymorphic in African HapMap populations. No HapMap data was available for rs11066015, but this SNP is in strong linkage disequilibrium (LD) with rs671 in the Chinese population, which we found to be absent in both the Black and Mixed Ancestry South African populations (10). Thus, these three SNPs are likely to be rare or absent in our South African populations, and we would have limited power to detect association with OSCC. The five prioritized SNPs were genotyped in 407 cases and 849 controls from the South African Black population and 257 OSCC cases and 860 controls from the Mixed Ancestry population using validated TaqMan 5′ exonuclease SNP genotyping assays (Applied Biosystems). Reactions were carried out in 2.5 µl volumes in 96-well plates. Each reaction contained 20 ng DNA, Absolute QPCR ROX mix (ABgene) and TaqMan SNP assay mix (Applied Biosystems) according to assay instructions and were performed on a PTC-0225 DNA Engine (MJ Research). Fluorescent levels at the PCR endpoint were determined using a 7900HT Fast Real-Time PCR system (Applied Biosystems) and genotypes assigned using SDS 2.2.2 software (Applied Biosystems).

Sequencing of PLCE1

The 34 exons of PLCE1 were sequenced by Sanger sequencing in 46 cancer patients (38 OSCC and 8 other cancers) from the Black South African population. Exons were amplified by PCR with primers designed using Primer3 (15) and synthesized by Integrated DNA Technologies. PCR was carried out in a 10 µl reaction containing 10 ng DNA, 5 µl 2× PCR mix (Promega), 0.4 pmoles of each forward and reverse primer for all exons apart from exon 1. For this exon, the reaction contained 1× Flexi reaction buffer (Promega), 1.5 mM $MgCl_2$, 0.4 pmoles of each forward and reverse primer, 1 U Flexi *Taq* polymerase (Promega), 0.2mM dNTP and 10% dimethyl sulfoxide. All reactions were carried out on a PTC-0225 DNA Engine (MJ Research) using the following conditions: 2 min at 92°C; then 30 cycles of 20 s at 92°C, 30 s at the optimized annealing temperature, and between 30 s and 3 min at 72°C (depending on amplimer length); with a final 5 min at 72°C. Primer sequences and PCR conditions for each amplicon are shown in Supplementary Table 1, available at *Carcinogenesis* Online. Subsequent ExoSAP-IT clean up (USB Europe, Staufen, Germany) followed by forward and/or reverse cycle sequencing was performed for each exon using 8 pmoles of sequencing primer (see Supplementary Table 2, available at *Carcinogenesis* Online) and 0.25 µl of BigDye Terminator v3.1 (Applied Biosystems) in a 5.25 µl reaction volume under recommended reaction conditions. Products were analysed on an ABI3730xl DNA sequencer (Sequence Analysis, Applied Biosystems) and aligned to the human reference genome using Staden software package (16).

Genotyping of PLCE1 SNPs

Three non-synonymous SNPs in *PLCE1*, Arg548Leu (rs17417407), Pro1890Leu (rs58539480), and the novel variant Gly1199Ser were selected for genotyping in the Black South African population on the basis of strong evolutionary conservation of the amino acid residue, a predicted damaging effect in at least one of the two programs, Polyphen 2 (17) and SIFT (Sorting Intolerant From Tolerant) (18), and a minor allele frequency (MAF) of >0.05 in this population. The variant Ile1777Thr (rs3765524) was also selected since it was very strongly associated with OSCC in the Chinese population and common in the Black South African population. A common insertion/deletion (indel) in the 5′-untranslated region (UTR) was also selected. These five variants were genotyped in an initial sample of 323 cases and 459 controls that was available from the Black South African population. If a suggestive association with OSCC was observed ($P < 0.05$) then further samples that subsequently became available were genotyped in an expanded sample set from this population (a total of 407 cases and 849 controls), and in the Mixed Ancestry population (257 cases and 860 controls). The SNPs were genotyped using custom KASP By-Design assays (KBioscience) following manufacturer's instructions. The 14 bp indel (CCCGGGCTCTGCCT) in the 5′UTR of exon 1 was PCR amplified and genotyped by size separation of PCR products using 3% agarose gel electrophoresis and visualized with ethidium bromide/UV light. Primers used for amplification were as follows: forward GGGAGCGGACTGTGAACG and reverse GTGTCCCCGCTACTGTGTGT. The 10 µl PCR reaction contained 1× Flexi reaction buffer (Promega), 1.5 mM $MgCl_2$, 0.4 pmoles of each forward and reverse primer, 1 U Flexi *Taq* polymerase (Promega), 0.2 mM dNTP and 10% DMSO. Reaction conditions were as described previously, with 63°C annealing temperature and 30 s extension time.

Statistical analysis

Pearson's chi-squared (χ^2) test was used to determine whether the proportions of genotypes were consistent with Hardy–Weinberg equilibrium, using a cut-off of $P < 0.05$. P values were >0.05 for all SNPs tested. Genotype and allele frequencies were calculated for cases and controls and the allele frequencies compared using the Pearson's chi-squared (χ^2) test to test for association with OSCC. For the association tests of the five Chinese GWAS SNPs, a Bonferroni-corrected P value of <0.01 (0.05/5) was used as a significance threshold to account for multiple testing. This threshold was also applied to the association tests of the five *PLCE1* variants genotyped after identification by sequencing. No additional correction was applied for the two populations tested. Allelic odds ratios (OR) and 95% confidence intervals (CI) were calculated using the common allele as the reference. As a secondary analysis, logistic regression was carried out adjusting for age, sex, smoking and alcohol consumption to determine whether these covariates influenced the association results determined using the Pearson's (χ^2) test; these results are reported as $P_{(adjusted)}$. For SNPs with suggestive evidence of allelic association ($P < 0.05$), the effect of alcohol and tobacco was investigated by testing for association in cases and controls stratified by smoking and drinking status. We tested for interactions by performing a case-only analysis of alleles based on smoking and drinking status, and by carrying out a gene–environment interaction test using logistic regression in a case–control analysis. The power of the study was determined using Quanto (http://hydra.usc.edu/gxe/). LD between *PLCE1* variants in the South African Black population was assessed using Haploview (19). *PLCE1* haplotype analysis was performed using UNPHASED (20).

Table I. Characteristics of OSCC cases and controls in the South African Black and Mixed Ancestry populations

		Black population		Mixed Ancestry population	
		Cases	Controls	Cases	Controls
n		407	849	257	860
*Age, mean years (SD)		59.8 (11.3)	48.8 (16.7)	60.6 (10.6)	46.7 (16.8)
*Sex, *n* (%)	Male	199 (48.9%)	335 (39.5%)	165 (64.2%)	309 (35.9%)
	Female	208 (51.1%)	511 (60.2%)	91 (35.4%)	551 (64.1%)
	Unknown	0 (0.0%)	3 (0.4%)	1 (0.4%)	0 (0.0%)
*Smoking status, *n* (%)	Ever-smoker	242 (59.5%)	333 (39.2%)	240 (93.4%)	597 (69.4%)
	Never-smoker	164 (40.3%)	505 (59.5%)	15 (5.8%)	258 (30.0%)
	Unknown	1 (0.2%)	11 (1.3%)	2 (0.8%)	5 (0.6%)
*Alcohol consumption, *n* (%)	Drinker	253 (62.2%)	452 (53.2%)	212 (82.5%)	419 (48.7%)
	Non-drinker	151 (37.1%)	393 (46.3%)	45 (17.5%)	436 (50.7%)
	Unknown	3 (0.7%)	4 (0.5%)	0 (0.0%)	5 (0.6%)

*Age, sex, smoking status and alcohol consumption were significantly different between cases and controls in both populations ($P < 0.01$).

Results

Case–control analysis

Five SNPs (*PLCE1* rs2274223, *RUNX1* rs2014300, *C20orf54* rs13042395, *PDE4D* rs10052657 and a variant near *UNC5CL* rs10484761) were tested for association in the South African Black and Mixed Ancestry populations (Table II). In the South African Mixed Ancestry population, the minor 'A' allele of rs2014300 in *RUNX1* was significantly associated with an increased risk of OSCC (OR = 1.33, 95% CI = 1.09–1.63, $P = 0.0055$), with minor allele frequencies of 43.8% and 37.0% in cases and controls, respectively. However, this effect is in the opposite direction to that found in the Chinese population, where the minor 'A' allele confers a reduced risk of OSCC (OR = 0.70) (14). The other four SNPs were not associated with OSCC in this population. In the Black South African population there was no evidence of association with OSCC for any of the variants tested. In both populations, adjusting for covariates in logistic regression produced very similar results. None of the five loci was associated with OSCC in the Black population, and ORs were similar to the unadjusted analysis. In the Mixed Ancestry population, the four loci not associated with the unadjusted analysis were also not associated with the adjusted analysis, and the *RUNX1* variant rs201400 remained associated with a moderately increased effect ($OR_{adjusted} = 1.51$, 95% CI = 1.19–1.92; $P_{(adjusted)} = 0.0007$). There were substantial differences in allele frequencies for three of the five SNPs in the two South African populations; for rs2014300 and rs10484761, the minor allele in the Black population was the common allele in the Mixed Ancestry population, and the 'T' allele of rs13042395 was very rare (0.5%) in the Black population compared with the Mixed Ancestry population (6.8%) (Table II).

The association tests in the South African populations had good power to detect the ORs reported in the Chinese GWAS studies for four out of the five loci tested, if it is assumed that the associated variant and the causal variant are either the same or are in complete LD. The ORs reported in the Chinese studies were 1.34–1.43 (*PLCE1*), 1.33 (*UNC5CL*), 0.67 (*PDE4D*), 0.70 (*RUNX1*) and 0.66 (*C20orf54*) (12–14). Using the allele frequencies determined in the Black controls, power in the Black population was >80% to detect the effects seen in *PLCE1*, *UNC5CL*, *PDE4D* and *RUNX1*, and there was 80% power to detect ORs of 1.27, 1.28, 1.49 and 1.28 for these four loci, respectively. In the Mixed Ancestry population, power was somewhat lower but still adequate at >75% for these four SNPs, with 80% power to detect ORs of 1.33, 1.35, 1.54 and 1.35, respectively. The low frequency of the *C20orf54* SNP rs13042395 in both populations (0.5% and 6.8% in Black and Mixed Ancestry controls, respectively) did not provide sufficient power to detect the association reported in the Chinese population.

Sequencing and genotyping of PLCE1

The strong association of the SNP rs2274223 (His1927Arg) in *PLCE1* with OSCC in the Chinese population (12–14) was not replicated in either of the South African populations tested. Since genetic variability is highest in African populations and LD is generally much lower, we investigated the *PLCE1* locus in the Black South African population by sequencing all 34 exons of *PLCE1* and adjacent splice sites in 46 individuals from this population to examine LD and to identify potential functional sequence variants. A total of 48 polymorphic variants were detected, including 26 known SNPs, 11 novel variants and one known 14 bp insertion/deletion in the 5′-UTR. Ten of these 48 variants produce amino acid substitutions in the encoded PLCE1 protein and could therefore be considered as candidate OSCC causal variants; their locations relative to the putative protein domains of PLCE1 are shown in Figure 1. The potential functional consequences of these variants were assessed by examination of evolutionary conservation across multiple species and use of the predictive programs Polyphen2 and SIFT (see Materials and methods). Six of the ten SNPs were predicted to have probable or possible damaging effects by one or more of these methods (Table III).

Pair-wise analysis of all 48 variants detected by sequencing in the 46 Black South African individuals showed that there is very low LD across the *PLCE1* gene in this population (Supplementary Figure 1, available at *Carcinogenesis* Online). Sixteen of these variants have also been genotyped in the HapMap project, allowing a comparison of the LD structure between the South African Black population and HapMap populations such as the Han Chinese from Beijing (CHB) and Yoruba in Ibadan, Nigeria (YRI) (Figure 2; Supplementary Figure 2,

Table II. Case–control analysis of five loci from Chinese GWAS in South African Black and Mixed Ancestry OSCC

	Black population					Mixed Ancestry population				
		Cases	Controls	OR (95% CI)	*P* value		Cases	Controls	OR (95% CI)	*P* value
PLCE1	His/His	140 (33.5%)	302 (35.5%)			His/His	78 (30.7%)	310 (36.2%)		
rs2274223	His/Arg	208 (49.8%)	411 (48.4%)			His/Arg	130 (51.2%)	408 (47.6%)		
	Arg/Arg	70 (16.7%)	137 (16.1%)			Arg/Arg	46 (18.1%)	139 (16.2%)		
	His	488 (58.4%)	1015 (59.7%)	Reference	—	His	286 (56.3%)	1028 (60.0%)	Reference	—
	Arg	348 (41.6%)	685 (40.3%)	1.06 (0.89–1.25)	0.5208	Arg	222 (43.7%)	686 (40.0%)	1.16 (0.95–1.42)	0.1386
C20orf54	C/C	402 (99.5%)	837 (98.9%)			C / C	221 (87.0%)	742 (87.2%)		
rs13042395	C/T	2 (0.5%)	9 (1.1%)			C / T	32 (12.6%)	103 (12.1%)		
	T/T	0 (0.0%)	0 (0.0%)			T / T	1 (0.4%)	6 (0.7%)		
	C	806 (99.8%)	1683 (99.5%)	Reference	—	C	474 (93.3%)	1587 (93.2%)	Reference	—
	T	2 (0.2%)	9 (0.5%)	0.46 (0.10–2.15)	0.3150	T	34 (6.7%)	115 (6.8%)	0.99 (0.67–1.47)	1.0000
near *UNC5CL*	G/G	107 (26.9%)	225 (26.6%)			A / A	105 (41.5%)	398 (47.2%)		
rs10484761	G/A	210 (52.8%)	435 (51.4%)			G / A	117 (46.2%)	362 (42.9%)		
	A/A	81 (20.4%)	186 (22.0%)			G / G	31 (12.3%)	84 (10.0%)		
	G	424 (53.3%)	885 (52.3%)	Reference	—	A	327 (64.6%)	1158 (68.6%)	Reference	—
	A	372 (46.7%)	807 (47.7%)	0.96 (0.81–1.14)	0.6542	G	179 (35.4%)	530 (31.4%)	1.20 (0.97–1.47)	0.0933
PDE4D	C/C	300 (74.6%)	642 (76.0%)			C / C	171 (67.1%)	613 (71.6%)		
rs10052657	C/A	94 (23.4%)	190 (22.5%)			C / A	79 (31.0%)	220 (25.7%)		
	A/A	8 (2.0%)	13 (1.5%)			A / A	5 (2.0%)	23 (2.7%)		
	C	694 (86.3%)	1474 (87.2%)	Reference	—	C	421 (82.5%)	1446 (84.5%)	Reference	—
	A	110 (13.7%)	216 (12.8%)	1.08 (0.85–1.38)	0.5329	A	89 (17.5%)	266 (15.5%)	1.15 (0.88–1.50)	0.3005
RUNX1	A/A	152 (37.7%)	311 (37.1%)			G / G	81 (32.1%)	346 (40.9%)		
rs2014300	A/G	197 (48.9%)	378 (45.1%)			A / G	121 (48.0%)	374 (44.2%)		
	G/G	54 (13.4%)	149 (17.8%)			A / A	50 (19.8%)	126 (14.9%)		
	A	501 (62.2%)	1000 (59.7%)	Reference	—	G	283 (56.2%)	1066 (63.0%)	Reference	—
	G	305 (37.8%)	676 (40.3%)	0.90 (0.76–1.07)	0.2342	A	221 (43.8%)	626 (37.0%)	1.33 (1.09–1.63)	0.0055

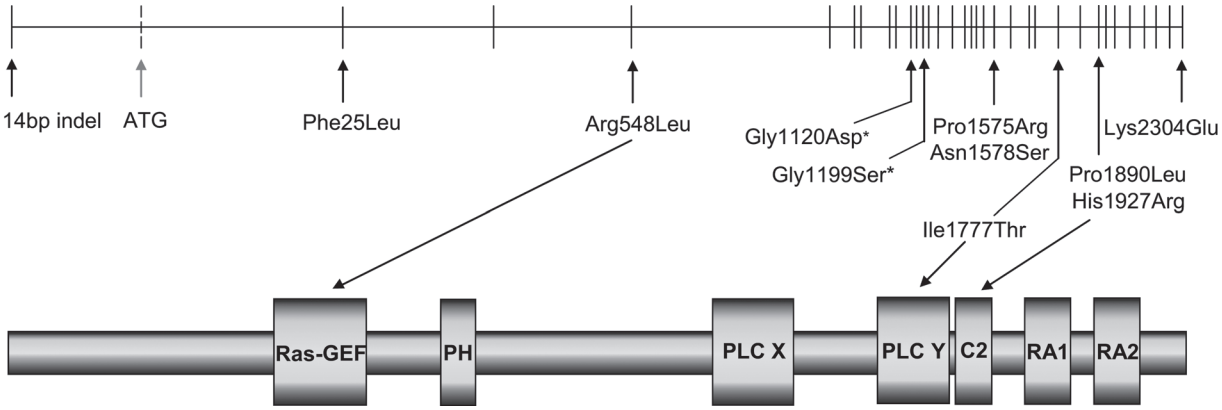


Fig. 1. Potential functional variants detected in the *PLCE1* gene. Top shows exon/intron structure with vertical lines representing exons for NM_016341.3 (dotted vertical line reflects exon present only in NM_001165979.1). Locations of non-synonymous SNPs and 5′-UTR insertion/deletion identified by sequencing in the South African Black population are indicated with vertical arrows (* indicates novel variants). Bottom shows the predicted functional domains of the PLCE1 protein and the position of variants which are located within a domain.

available at *Carcinogenesis* Online). This shows that the South African Black population has the lowest level of LD across *PLCE1* among these three populations. The index SNP rs2274223 is in very strong LD with multiple other SNPs in the CHB population, but is in low to moderate LD with other SNPs in the South Africans. The variants Arg548Leu (rs17417407), Ile1777Thr (rs3765524), Pro1890Leu (rs58539480) and the novel Gly1199Ser SNP, together with the 5′-UTR 14 bp indel, were selected for genotyping (see Materials and methods) and tested for association with OSCC in the Black South African population. Results for allelic association tests are shown in Table IV. Only one variant, *PLCE1* Arg548Leu (rs17417407), showed suggestive evidence for association with OSCC (P = 0.035) and was therefore genotyped in an additional set of cases and controls. Combined analysis of 407 cases and 849 controls revealed a MAF (T, 548Leu) of 16.6% in cases and 21.1% in controls, giving an OR = 0.74 (95% CI = 0.60–0.93, P = 0.008). Similar results were obtained for these five *PLCE1* variants when adjusting for age, sex, smoking and drinking status, except that the rs1741707 association was slightly less significant ($OR_{adjusted}$ = 0.75, 95% CI = 0.59–0.95, $P_{(adjusted)}$ = 0.019). Haplotype analysis of the five SNPs and indel of *PLCE1* show that none of the haplotypes were associated with OSCC in the Black South African population (P = 0.977 for overall association, data not shown). The Arg548Leu SNP (rs17417407) was also genotyped in the Mixed Ancestry population, but no evidence of association was detected, with MAF for the Leu allele of 17.4% and 18.0% in cases and controls, respectively (OR = 0.96, 95% CI = 0.74–1.25, P = 0.764).

Alcohol and smoking analysis

Gene–environment interactions between OSCC and smoking and alcohol drinking habits were tested for variants that showed an association with OSCC ($P < 0.05$). In the Mixed Ancestry population, *RUNX1* rs2014300 showed no significant associations in case–control analysis for either drinkers or non-drinkers, or for a case-only analysis of drinkers versus non-drinkers (see Supplementary Table 3, available at *Carcinogenesis* Online). The low number of non-smokers (n = 15) in the Mixed Ancestry population prevented analysis by smoking status. The *PLCE1* Arg548Leu SNP (rs17417407) showed no interaction with alcohol use in the Black population. When stratifying by smoking status in cases and controls, an association was observed in a case–control analysis of ever-smokers, OR = 0.64 (P = 0.005), with weakened association in never-smokers (OR = 0.87). However, the case-only analysis of ever- versus never-smokers was not significant (P = 0.16). Logistic regression analysis showed no evidence for an interaction with smoking or alcohol with either *RUNX1* rs2014300 or *PLCE1* Arg548Leu (rs17417407).

Discussion

The aim of this study was to determine whether five new loci reported to be associated with susceptibility to OSCC in the Chinese population (12–14) also contribute to susceptibility in two South African populations. In the South African Mixed Ancestry population, only one SNP, *RUNX1* rs2014300, was associated with OSCC, with the minor

Table III. *PLCE1* non-synonymous SNPs in the South African Black population

SNP identifier	Chr location (build 37), major > minor allele	Amino acid change	MAF[a]	Amino acid conservation[b]	Polyphen 2 (score)	SIFT (score)
rs115135156	10:95848924, T>C	Phe25Leu	0.043	Rh[c], Mo[c], Do[c], El[c], Op[c]	Benign (0.013)	Damaging (0)[d]
rs17417407	10:95931087, G>T	Arg548Leu	0.178	Rh, Mo, Do, El, Op, Ch, X_t, Ze	Probably damaging (0.981)	Tolerated (0.12)
Novel	10:96014026, G>A	Gly1120Asp	0.011	Rh, Mo, Do, El, Op, Ch, X_t	Probably damaging (0.968)	Damaging (0.04)
Novel	10:96018597, G>A	Gly1199Ser	0.058	Rh, Mo, Do, El, Op, Ch, X_t	Probably damaging (0.984)	Tolerated (0.12)
rs2274224	10:96039597, C>G	Pro1575Arg	0.318	Not conserved	Benign (0)	Tolerated (0.53)
rs61732525	10:96039606, A>G	Asn1578Ser	0.045	Rh, Do, El, X_t, Ze	Benign (0.002)	Tolerated (0.42)
rs3765524	10:96058298, T>C	Ile1777Thr	0.386	Not conserved	Benign (0.002)	Tolerated (0.85)
rs58539480	10:96066230, C>T	Pro1890Leu	0.087	Rh, Mo, Do, El, Op, X_t, Ze	Probably damaging (0.999)	Tolerated (0.33)
rs2274223	10:96066341, A>G	His1927Arg	0.489	Ze	Benign (0)	Tolerated (0.83)
rs3203713	10:96087728, A>G	Lys2304Glu	0.141	Rh[c], Do[c], El[c]	Unknown	Damaging (0.01)[d]

[a]MAF calculated from the number of individuals genotyped successfully (n = 43–46).
[b]Amino acid conservation is shown for Rhesus Macaque (Rh), Mouse (Mo), Dog (Do), Elephant (El), Opossum (Op), Chicken (Ch), Xenopus Tropicalis (X_t) and Zebrafish (Ze).
[c]Conservation only available for nucleotide sequence.
[d]Low confidence SIFT score indicating that the protein alignment does not have enough sequence diversity and an amino acid may incorrectly be predicted to be damaging.

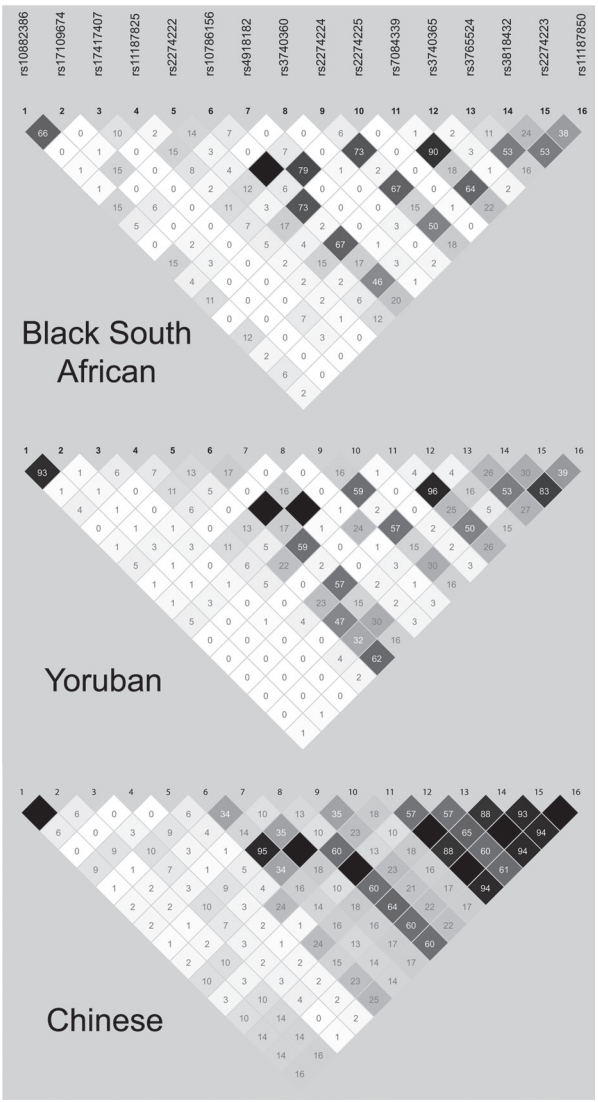


Fig. 2. HapMap LD structure for 16 SNPs across *PLCE1* gene region. Plots show r^2 values for pairwise LD between variants in the South African Black population and two HapMap populations, Yoruban from Nigeria and Han Chinese from Beijing. Colour schemes and labelling are according to Haploview as follows: white ($r^2 = 0$), black ($r^2 = 1$) or shades of grey ($0 < r^2 < 1$); numbers indicate r^2 values (×100).

'A' allele conferring an increased risk of disease (OR = 1.33, 95% CI = 1.09–1.63). However, the association is in the opposite direction to that found in the Chinese population, where allele 'A' is also the minor allele but is protective (OR = 0.70) (14). Since it is unlikely that the same allele of this SNP would have opposite effects on susceptibility to the same disease in these two populations, the finding in the Mixed Ancestry population may be a false positive, despite remaining significant after correction for multiple testing. Further analysis of this SNP in other populations would help to establish whether this association is specific to OSCC in the Chinese population.

In the South African Black population, none of the five SNPs showed evidence of association with OSCC. This is consistent with our previous study, in which none of the 13 variants associated with OSCC in other populations was associated in South African Black OSCC (10). There are several possible reasons for the lack of association of the five SNPs in the South African populations. One is that the GWAS findings are false positives. This is very unlikely for *PLCE1* since the association has been observed in three independent studies (12–14). The associations at *PDE4D* and *RUNX1* were convincingly replicated in the original study (14), with combined *P* values of 10^{-19} and 10^{-21}, respectively, so these are also probably to be robust findings. Replication of the *UNC5CL* locus was less convincing (14), and the *C20orf54* association was only observed in one of the three GWAS (13). A meta-analysis of all three Chinese GWAS would help to resolve the status of some of these loci. An alternative explanation for the lack of association in South African OSCC is insufficient power. We had high power to detect the effect observed in the Chinese studies for four of the five SNPs, with the other SNP (*C20orf54* rs13042395) being very rare in the Black population. However, if the effect size at these loci is much smaller in South Africans than in Chinese, the power of this study would be reduced. A further possible reason is that the SNPs genotyped in the Chinese GWAS studies may not be the actual causal SNPs that are driving the association, but merely markers that tag them with high LD. Since LD is generally lower in African populations (21), our studies may not be able to detect an association if the causal SNP is not genotyped directly.

Table IV. Case–control analysis of five potential functional *PLCE1* variants for OSCC in the South African Black population

		Cases	Controls	OR (95% CI)	*P* value
5′-UTR 14 bp indel	Ins/Ins	185 (57.6%)	260 (57.0%)		
	Ins/Del	122 (38.0)	171 (37.5%)		
	Del/Del	14 (4.4%)	25 (5.5%)		
	Ins	492 (76.6%)	691 (75.8%)	Reference	—
	Del	150 (23.4%)	221 (24.2%)	0.95 (0.75–1.21)	0.693
rs17417407 Arg548Leu	Arg/Arg	226 (70.4%)	271 (61.5%)		
	Arg/Leu	81 (25.2%)	152 (34.5%)		
	Leu/Leu	14 (4.4)	18 (4.1%)		
	Arg	533 (83.0%)	694 (78.7%)	Reference	—
	Leu	109 (17.0%)	188 (21.3%)	0.75 (0.58–0.98)	0.035
Novel Gly1199Ser	Gly/Gly	289 (90.0%)	410 (91.3%)		
	Gly/Ser	30 (9.3%)	38 (8.5%)		
	Ser/Ser	2 (0.6%)	1 (0.2%)		
	Gly	608 (94.7%)	858 (95.5%)	Reference	—
	Ser	34 (5.3%)	40 (4.5%)	1.20 (0.75–1.92)	0.446
rs3765525 Ile1777Thr	Ile/Ile	86 (27.2%)	126 (27.9%)		
	Ile/Thr	162 (51.3%)	233 (51.5%)		
	Thr/Thr	68 (21.5%)	93 (20.6%)		
	Ile	334 (52.8%)	485 (53.7%)	Reference	—
	Thr	298 (47.2%)	419 (46.3%)	1.03 (0.84–1.27)	0.756
rs58539480 Pro1890Leu	Pro/Pro	262 (85.3%)	375 (87.4%)		
	Pro/Leu	45 (14.7%)	53 (12.4%)		
	Leu/Leu	0 (0.0%)	1 (0.2%)		
	Pro	569 (92.7%)	803 (93.6%)	Reference	—
	Leu	45 (7.3%)	55 (6.4%)	1.15 (0.77–1.74)	0.490

The importance of the *PLCE1* locus in OSCC susceptibility in the Chinese population prompted us to investigate sequence variation in this gene and its genetic architecture in the South African Black population in more detail, since substantial differences in sequence variation and LD structure could exist between these two populations. Sequencing of the entire coding region in 46 individuals revealed that LD across *PLCE1* was much weaker in the South African Black population compared with the Chinese. This suggests that very high density SNP analysis across its entire genomic structure would be required for a complete interrogation of the contribution of *PLCE1* to OSCC in this population. We tested five coding sequence changes that were conserved and/or predicted to alter protein function and an insertion/deletion polymorphism in the 5′-UTR as these were sufficiently common to allow detection of association with OSCC in the South African Black population. Only the Arg548Leu variant (rs17417407) was associated with OSCC. The Arg548 allele is conserved across species (Rhesus, Mouse, Dog, Elephant, Opossum, Chicken, Xenopus Tropicalis and Zebrafish) and is predicted to

be 'probably damaging' by Polyphen2 but tolerated by SIFT. This SNP was not included in the GWAS SNP micorarrays used by the three Chinese studies. However, it is in complete LD (r^2 = 1) with rs2689700 in the combined Chinese/Japanese (CHB/JPT) HapMap population, which is an intronic SNP that was present on the GWAS chips used in all three GWAS studies and is not listed as associated with OSCC. In both the Chinese/Japanese and South African Black populations, Arg548Leu rs17417407 is in very low LD with His1927Arg rs2274223 (r^2 = 0.023 and 0.03, respectively), the Chinese OSCC index risk variant. This suggests that two independent variants in *PLCE1* may contribute to OSCC susceptibility loci in the Chinese and South African Black populations.

An important question regarding the association at the *PLCE1* locus is whether variants within *PLCE1* itself are driving this association, since the GWAS signal appears to include at least one other adjacent gene, *NOC3L* (12–14). The protein encoded by *PLCE1*, phospholipase C epsilon 1, is responsible for the hydrolysis of phosphatidyl-inositol 4,5-biphosphate to generate diacylglycerol and inositol 1,4,5-triphosphate, which causes the release of calcium and activation of protein kinase C. PLCE1 can also act as a guanine-exchange factor, activating Ras, which is unique to this class of phospholipase C enzymes; PLCE1 is itself activated by Ras and Rho family GTPases, and thus could be affected by the oncogenic properties of Ras in cancer cells, although this has not yet been shown (reviewed in ref. 22). Studies of *PLCE1* expression and protein levels in tumours have produced conflicting results, with mRNA shown to be reduced in OSCC tissue (23), but with unchanged (23) or increased levels (13) of the protein being reported in tumour tissues compared with normal tissue. Activation of PLCE1 has also been linked to tumour cell migration in head and neck squamous cell carcinoma (24). However, sequence variation in *PLCE1* has also been associated with non-cancer phenotypes. The Ile1777Thr variant (rs3765524), which is strongly associated with OSCC (12), is also associated with dengue shock syndrome (25), and biallelic mutations in *PLCE1* are an important cause of nephrotic syndrome (26). The adjacent gene, *NOC3L* (also known as *FAD24*) has been shown to regulate DNA replication during adipogenesis (27). It has also been reported to have a role in repression of NF-κB activity and H-Ras-mediated transformation (28), and silencing of the gene produces sensitivity to Tamoxifen, a drug that inhibits estrogen receptor α signalling in breast cancer (29). Fine mapping of the region of association with a dense panel of SNPs in parallel with functional are needed to confirm the identity of the causal gene.

Two of the SNPs showing the strongest association at the *PLCE1* locus in the Chinese population are the non-synonymous variants Ile1777Thr (rs3765524) and His1927Arg (rs2274223), which raises the question of whether one or other of them might be the causal variant at this locus. His1927Arg is located within the protein calcium-dependent lipid-binding C2 domain of PLCE1, but our bioinformatic analysis of this variant (Table III) shows that it is not conserved and is scored as benign or tolerated by the functional prediction programs PolyPhen and SIFT. Ile1777Thr is located in the PI-PLC Y-box domain that is important for catalytic activity, but is not evolutionary conserved and is also scored as benign or tolerated by the prediction programs. Thus, neither variant has strong credentials as the causal variant. Interestingly, the Arg548Leu variant rs17417407, which showed modest association with OSCC in the South African Black population, is located in a Ras-guanine-exchange factor domain, and may therefore affects its capacity to activate the Ras protein. It does show strong evolutionary conservation and is predicted to be damaging by one of the two prediction programs. It should also be noted that non-coding variants and synonymous SNPs may have important functional effects (30,31), and multiple associations with complex disease have been detected in gene deserts, which do not contain any known coding genes (32). These classes of variants will need to be taken into account when attempting to define causal variants at complex disease loci.

In conclusion, this study has examined a series of new genetic associations with OSCC, which have emerged from GWAS in the Chinese population, in two indigenous populations from South Africa. None of these associations was replicated in either of the two indigenous South African populations studied. However, although none of the variants in *PLCE1* that conferred an increased risk of OSCC in Chinese populations were associated in the South African populations, we found that the Arg548Leu variant rs17417407 is associated with a reduced risk of OSCC in the Black South African population. As discussed previously (10), the reasons for variation in genetic associations between populations may include differences in both genetic architecture and in environmental risk factors that interact with genetic factors to promote development of the disease. The emerging differences that we have observed between the genetic factors contributing to the development of OSCC in African versus Asian populations (9,10) suggest that well-powered GWAS in non-Asian populations are needed to define the genetic basis of this cancer in Africa and elsewhere.

Supplementary material

Supplementary Tables 1–3 and Figures 1 and 2 can be found at http://carcin.oxfordjournals.org/

Funding

Association for International Cancer Research (09-0625), the Medical Research Council UK, The Generation Trust, the National Institutes of Health Research Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London to H.B., N.J.P., C.M.L., C.G.M., The South African Research Chairs Initiative of the Department of Science and Technology and the National Research Foundation, the International Centre for Genetic Engineering and Biotechnology (ICGEB), the South African Medical Research Council and the University of Cape Town to M.I.P.

Acknowledgements

We thank Antoinette Olivier, Zenaria Abbas and Amy Salkinder for assisting with the sample collection and processing, and the patients and healthy controls for their participation in this study.

Conflict of Interest Statement: None declared.

References

1. Ferlay, J. *et al.* (2010) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer*, **127**, 2893–2917.
2. Jemal, A. *et al.* (2012) Cancer burden in Africa and opportunities for prevention. *Cancer*, doi: 10.1002/cncr.27410. [Epub ahead of print.]
3. Somdyala, N.I. *et al.* (2010) Cancer incidence in a rural population of South Africa, 1998–2002. *Int. J. Cancer*, **127**, 2420–2429.
4. Hendricks, D. *et al.* (2002) Oesophageal cancer in Africa. *IUBMB Life*, **53**, 263–268.
5. Li, D. *et al.* (2010) The 341C/T polymorphism in the GSTP1 gene is associated with increased risk of oesophageal cancer. *BMC Genet.*, **11**, 47.
6. Dandara, C. *et al.* (2005) CYP3A5 genotypes and risk of oesophageal cancer in two South African populations. *Cancer Lett.*, **225**, 275–282.
7. Li, D. *et al.* (2005) Association of cytochrome P450 2E1 genetic polymorphisms with squamous cell carcinoma of the oesophagus. *Clin. Chem. Lab. Med.*, **43**, 370–375.
8. de Wit, E. *et al.* (2010) Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Hum. Genet.*, **128**, 145–153.
9. Matejcic, M. *et al.* (2011) Association of a deletion of GSTT2B with an altered risk of oesophageal squamous cell carcinoma in a South African population: a case-control study. *PLoS ONE*, **6**, e29366.
10. Bye, H. *et al.* (2011) Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa. *Carcinogenesis*, **32**, 1855–1861.
11. Cui, R. *et al.* (2009) Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. *Gastroenterology*, **137**, 1768–1775.

12. Abnet, C.C. *et al.* (2010) A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nat. Genet.*, **42**, 764–767.
13. Wang, L.D. *et al.* (2010) Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies susceptibility loci at PLCE1 and C20orf54. *Nat. Genet.*, **42**, 759–763.
14. Wu, C. *et al.* (2011) Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nat. Genet.*, **43**, 679–684.
15. Rozen, S. *et al.* (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
16. Bonfield, J.K. *et al.* (1995) A new DNA sequence assembly program. *Nucleic Acids Res.*, **23**, 4992–4999.
17. Adzhubei, I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
18. Kumar, P. *et al.* (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
19. Barrett, J.C. *et al.* (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
20. Dudbridge, F. (2008) Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum. Hered.*, **66**, 87–98.
21. Teo, Y.Y. *et al.* (2010) Methodological challenges of genome-wide association analysis in Africa. *Nat. Rev. Genet.*, **11**, 149–160.
22. Bunney, T.D. *et al.* (2010) Phosphoinositide signalling in cancer: beyond PI3K and PTEN. *Nat. Rev. Cancer*, **10**, 342–352.
23. Hu, H. *et al.* (2012) Putatively functional PLCE1 variants and susceptibility to esophageal squamous cell carcinoma (ESCC): a case-control study in Eastern Chinese populations. *Ann. Surg. Oncol.*, **19**, 2403–2410.
24. Bourguignon, L.Y. *et al.* (2006) Hyaluronan-CD44 interaction with leukemia-associated RhoGEF and epidermal growth factor receptor promotes Rho/Ras co-activation, phospholipase C epsilon-Ca2+ signaling, and cytoskeleton modification in head and neck squamous cell carcinoma cells. *J. Biol. Chem.*, **281**, 14026–14040.
25. Khor, C.C. *et al.* (2011) Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. *Nat. Genet.*, **43**, 1139–1141.
26. Hinkes, B. *et al.* (2006) Positional cloning uncovers mutations in PLCE1 responsible for a nephrotic syndrome variant that may be reversible. *Nat. Genet.*, **38**, 1397–1405.
27. Johmura, Y. *et al.* (2008) FAD24, a regulator of adipogenesis, is required for the regulation of DNA replication in cell proliferation. *Biol. Pharm. Bull.*, **31**, 1092–1095.
28. Johmura, Y. *et al.* (2008) FAD24, a regulator of adipogenesis and DNA replication, inhibits H-RAS-mediated transformation by repressing NF-kappaB activity. *Biochem. Biophys. Res. Commun.*, **369**, 464–470.
29. Mendes-Pereira, A.M. *et al.* (2012) Genome-wide functional screen identifies a compendium of genes affecting sensitivity to tamoxifen. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 2730–2735.
30. Sauna, Z.E. *et al.* (2007) Silent polymorphisms speak: how they affect pharmacogenomics and the treatment of cancer. *Cancer Res.*, **67**, 9609–9612.
31. Cooper, G.M. *et al.* (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.*, **12**, 628–640.
32. Mathew, C.G. (2008) New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nat. Rev. Genet.*, **9**, 9–14.

Received March 30, 2012; revised July 17, 2012; accepted July 31, 2012

Supplementary Table I. PCR primers for amplification of *PLCE1* exons

Exon	Forward primer	Reverse primer	Annealing temp (°C)	Extension time	Product length (bp)
1	GGGAGCGGACTGTGAACG	GAGCGCGAGGACACTTTTC	63	30 sec	444
2	TTTTGGCTGGAAGCAGAAGT	GCTATCAATTGGAGTATCTGTTTTCA	57	2 min	1922
NM_001165979.1 Exon 1	CCTCCCTCGATTCTGGGTAT	AACAGTCCCCAGGATTCCAT	58	30 sec	472
3	TTTGCACTTGGAGCATCTGA	TTCTCTTAAAGAAGCGACTTTTACTT	55	30 sec	543
4	CAGAACTCTTCACTAAGCAGAGGA	CATGGAAGTGGCAAGACAGA	60	30 sec	568
5	CCCAGCCAGGACCTACAG	GCCTTTGGTGCTGAAAGAAG	60	30 sec	437
6	GGAATTTAGGCTCCTTGCTG	TCCAAGGATCTATGTGCTACCC	57	30 sec	482
7	TCCTGGAGGCTCTTGTTTTC	TTGGAATTGGTAAGGTTTGAAGA	57	30 sec	506
8	CACCTGGCCTCGGTTATTAG	TAAAAGCTGCCCAAGGTCAC	57	1 min	977
9, 10, 11	TGGGTGGCCAGATCATTATT	TTTGGAGAATCATGGCTTAGG	57	3 min	3108
12+13	AGCTTCAATCTTAAATAACTTGCAC	TTTCCCTACACAGCAGTAATAGC	55	30 sec	599
14	TCCTATCACTATGTGAAGCCAGAA	AGCCTGGCCACAGAGTAAGA	61	30 sec	591
15, 16	CAGCCTTCTTTTCTCATTCTCTTC	AGCCAGTTTTCCCACACATC	57	30 sec	722
17	CCATTTGCCCTTCTGCTTTA	GCTTGATGGTATGGGCTTGT	55	30 sec	459
18	TGGCCTTATCCTCATGCTTC	GTTGCAGTGAGCCAAGACTG	55	30 sec	425
19	GCTTCTTTCCTAGTTCCTCTTCC	TCTTGGGTGAGTGAGATGAGAG	57	30 sec	491
20	CATTGCATTTCGAGGGAATC	TGAATTCAGAAACTCCTGGACA	57	30 sec	433
21, 22	TGCTTCAAGCCATCATTTTG	TTCATGAGCATCAAGGCAAA	57	2 min	1506
23	GCATGCAGTTCTTGTTGCAT	CCAGCCTGAAATGCTGTTTT	58	30 sec	541
24	TCCAAGAGGTATTCTGATGTGG	AAACATCGGAGGCACAATTC	58	30 sec	592
25	TGGGACGAATGGGTGATTAT	TTCTGGGAACTAAATCTGTATGACC	58	30 sec	439
26	TCATTCACTTTGTCCATTCCAG	TGTGCTTCAAAAGTGCTCCA	58	30 sec	543
27	CTGTTGGTTGCATGCCTGT	GTGCCAAGTGTCAGCCATTA	58	30 sec	396
28	CAAATGGACTCTCATCTTTTGC	TCATCCACATGGACTTTTGC	57	30 sec	385
29	TGAATAAGTTGTGCCGTTGC	TGTGCAGAAGAATAAAACTGTTCA	57	30 sec	574
30	GCACAGTAGTTTCCTCCTCTCA	CACACACTCCCCTTTGAGGT	57	30 sec	484
31	TCTGGAAGATCCCCTTCATC	GACTGCTTAACCGCAAGCTC	57	30 sec	529
32	GAGCTTGCGGTTAAGCAGTC	CCATAGAGCCCTTGAAGAATG	57	1 min	1062
33	CATTGTGAGTACAGAGGAAACAGTC	TCTAGCCTGCCACCTGTTTT	57	1 min	692

Exon numbers are based on GenBank accession number NM_016341.3, apart from NM_001165979.1 exon 1, as indicated.

Supplementary Table II. Primers for Sanger sequencing of *PLCE1* exons

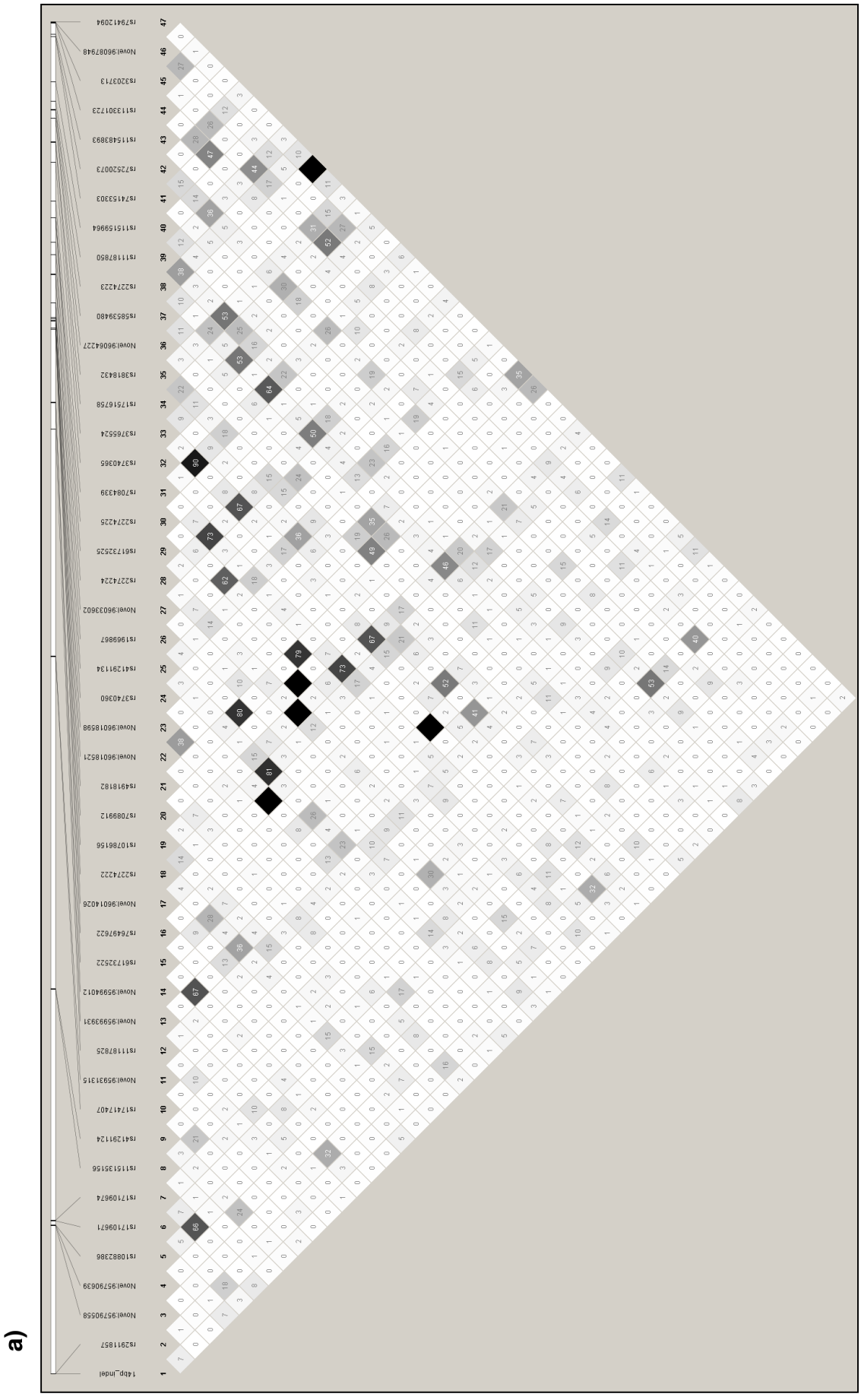
Exon	Forward primer	Reverse primer
1	GGGAGCGGACTGTGAACG	GAGCGCGAGGACACTTTTC
	TTTTGGCTGGAAGCAGAAGT	CACAGGTATGAGAACAGAAGCTG
	GATCTACCACCTTAAAACCCTGA	AGGAAGGCCATGCTGATG
	CACATACTGTCAGACGAAGTGG	
	CTGGAACTAGACAGACCTTCCA	
2	TGCTTTGAAGGCTCTTGTGA	
NM_001165979.1 Exon 1	CCTCCCTCGATTCTGGGTAT	AACAGTCCCCAGGATTCCAT
3	TTTGCACTTGGAGCATCTGA	
4	CAGAACTCTTCACTAAGCAGAGGA	
5	CCCAGCCAGGACCTACAG	
6	GGAATTTAGGCTCCTTGCTG	TCCAAGGATCTATGTGCTACCC
7	TCCTGGAGGCTCTTGTTTTC	
	CACCTGGCCTCGGTTATTAG	
	GACGGAGCTCATCCCTTG	
8	TGCTGGATTAAGTAGCCTGAC	
9	TGGGTGGCCAGATCATTATT	
10	CGGTCAGCCTTAATGTAGGTC	
11	CCACCAGATTAGCCCATTCA	
12+13	AGCTTCAATCTTAAATAACTTGCAC	TTTCCCTACACAGCAGTAATAGC
14	TCCTATCACTATGTGAAGCCAGAA	
15	CAGCCTTCTTTTCTCATTCTCTTC	ATGGCCCGTGAGGTAGGTAT
16	ATCCCTTGCAGAAGTTCGAG	AGCCAGTTTTCCCACACATC
17	CCATTTGCCCTTCTGCTTTA	
18	TGGCCTTATCCTCATGCTTC	
19	GCTTCTTTCCTAGTTCCTCTTCC	
20	CATTGCATTTCGAGGGAATC	
21	TGCTTCAAGCCATCATTTTG	
22	TCTAGGAAAGCTGTTGGGACA	
23	GCATGCAGTTCTTGTTGCAT	
24	TCCAAGAGGTATTCTGATGTGG	AAACATCGGAGGCACAATTC
25	TGGGACGAATGGGTGATTAT	
26	TCATTCACTTTGTCCATTCCAG	
27	CTGTTGGTTGCATGCCTGT	
28	CAAATGGACTCTCATCTTTTGC	
29	TGAATAAGTTGTGCCGTTGC	
30	GCACAGTAGTTTCCTCCTCTCA	
31	TCTGGAAGATCCCCTTCATC	
	GAGCTTGCGGTTAAGCAGTC	
	AAAGTATCCAAACCAAGGAGGA	
32	ATGCTATTGAACACCGCCTA	
	CATTGTGAGTACAGAGGAAACAGTC	
33	TTGATGATTTCTGAACTGAAGC	

Multiple primers were used for some exons to ensure complete coverage

Supplementary Table III. Stratified analysis of smoking and alcohol use for variants associated with OSCC

			Case-control: Drinkers only			Case-control: Non- drinkers only			Case-only: Drinkers vs. non-drinkers		
Variant	**Population**	**Minor allele**	**MAF: cases/ controls**	**OR (95% CI)**	**P-value**	**MAF: cases/ controls**	**OR (95% CI)**	**P-value**	**MAF: drinkers/ non-drinkers**	**OR (95% CI)**	**P-value**
RUNX1 rs2014300	Mixed Ancestry	A	0.454/0.403	1.23 (0.97 - 1.57)	0.0563	0.367/0.336	1.14 (0.73 - 1.79)	0.5642	0.454/0.367	1.44 (0.90 - 2.30)	0.1297
PLCE1 Arg548Leu	Black	T	0.174/0.220	0.73 (0.55 - 0.97)	0.0288	0.160/0.203	0.75 (0.53 - 1.07)	0.1111	0.171/0.160	1.09 (0.74 - 1.60)	0.6777
			Case-control: Ever smokers			**Case-control: Never smokers**			**Case-only: Ever vs. never**		
			MAF: cases/ controls	**OR (95% CI)**	**P-value**	**MAF: cases/ controls**	**OR (95% CI)**	**P-value**	**MAF: ever smokers/ never smokers**	**OR (95% CI)**	**P-value**
PLCE1 Arg548Leu	Black	T	0.150/0.216	0.64 (0.47 - 0.88)	0.0052	0.187/0.209	0.87 (0.63 - 1.20)	0.4003	0.150/0.187	0.77 (0.53 - 1.11)	0.1635

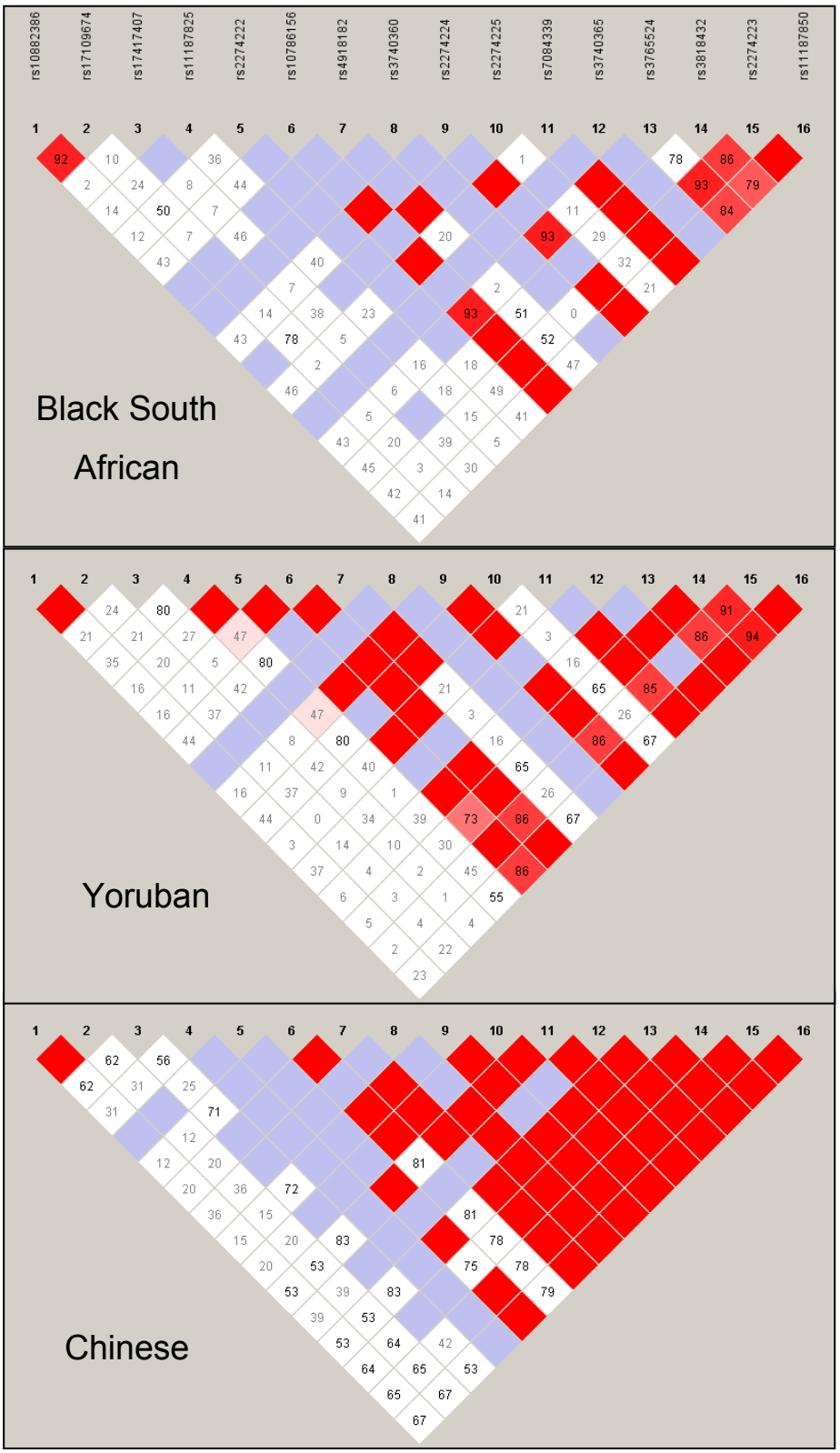
Smoking analysis could not be performed in the Mixed Ancestry population due to the low number of non-smokers (n=15)



a)

b)

Supplementary Fig. 1. Linkage disequilibrium (LD) between all 48 polymorphic variants in *PLCE1* identified by sequencing 46 individuals from the South African Black population; a) = r^2, b) = D'. Colour schemes and labelling are according to Haploview. For r^2 plots: white (r^2 = 0), black (r^2 = 1) or shades of grey (0 < r^2 < 1), and numbers indicate r^2 values (x 100). For D': white (D' < 1; LOD < 2), shades of pink/red (D' < 1; LOD ≥ 2), blue (D' = 1; LOD < 2) or bright red (D' = 1; LOD ≥ 2). Numbers indicate D' values (x 100). Where numbers are not shown, D' = 1.


rs10882386
rs17109674
rs17417407
rs11187825
rs2274222
rs10786156
rs4918182
rs3740360
rs2274224
rs2274225
rs7084339
rs3740365
rs3765524
rs3818432
rs2274223
rs11187850
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Black South African
Yoruban
Chinese

Supplementary Fig. 2. HapMap LD structure for 16 SNPs across *PLCE1* gene region. Plots show D' values for pairwise LD between variants in the South African Black population and two HapMap populations from Yoruban, Ibadan Nigeria and Han Chinese from Beijing. Colour schemes and labelling are according to Haploview: white (D' < 1; LOD < 2), shades of pink/red (D' < 1; LOD ≥ 2), blue (D' = 1; LOD < 2) or bright red (D' = 1; LOD ≥ 2). Numbers indicate D' values (x 100); where numbers are not shown, D' = 1.

4.4 Summary

This chapter investigated genetic susceptibility to OSCC in the South African Black and Mixed Ancestry populations, focusing on 5 loci that were associated with the disease in several GWAS in the Chinese population. The *RUNX1* SNP rs2014300 was associated with an increased risk of OSCC in the South African Mixed Ancestry population, but this effect was the opposite of that reported in the Chinese and thus is not a replication. None of the index SNPs at the 5 loci was associated with the disease in the Black population.

Since one of these variants, *PLCE1* His1927Arg (rs2274223), was associated with OSCC in all three independent Chinese GWAS, it provided strong evidence for the involvement of this gene in disease susceptibility in the Chinese population. Therefore, *PLCE1* was further investigated in the South African Black population by sequencing all exons to identify potential functional variants. Five potentially functional variants were selected for case-control association studies, and one of these, Arg548Leu (rs17417407), was found to be significantly associated with OSCC in the South African Black population.

Although the incidence of OSCC is high in both South African and Chinese populations, this work suggests that the genetic contribution to the disease differs between these populations.

5 Investigation of the genetic susceptibility to oesophageal cancer using the Immunochip

5.1 Involvement of the immune system in cancer development

A link between cancer and inflammation was first noted around 150 years ago, in 1863, when Rudolf Virchow observed that cancers originate at the site of chronic inflammation (reviewed in Balkwill and Mantovani 2001). Since that time, inflammation has been implicated in a number of different types of cancers, although the process is far from understood. Inflammation results in the recruitment of immune cells to a specific area which is controlled by a network of signalling molecules including cytokines, growth factors and chemokines. This inflammation may be caused by a number of factors, including chronic infections, chemical irritation and wound healing, which cause tissue damage.

A variety of cancers are thought to involve inflammation. For example, in lung cancer, inflammation is caused by tobacco smoke, a chemical irritant which also causes DNA mutations, enabling cells to survive in this environment. Oesophageal adenocarcinoma develops in some individuals who suffer from gastroesophageal reflux disease, which causes inflammation in the oesophagus (Barrett's oesophagus). World-wide, around 16% of cancers are thought to be attributable to infectious agents, with this figure rising to 32.7% in sub-Saharan Africa (de Martel *et al.* 2012). This includes cervical cancer (human papilloma virus), Burkitt's lymphoma (Epstein-Barr virus) and liver cancer (hepatitus viruses).

The mechanism by which inflammation is involved in cancer development is not fully understood. It is thought that the increased epithelial cell turnover needed to repair tissue damage brought about by inflammation may result in mutations in the replacement cells to allow them to survive in the harsh conditions (Moss and Blaser 2005). These cells have an abnormal growth, leading to metaplasia

and dysplasia, which is thought to be the precursor of cancer (Lu *et al.* 2006). Chronic activation of immune cells around these regions of pre-malignant tissues may promote tumour development by a number of mechanisms (reviewed in de Visser *et al.* 2006). This includes blocking CD8+ cytotoxic T lymphocyte responses to prevent the killing of tumour cells, and by promoting pro-tumour processes, including angiogenesis and proliferation, by the induction of signalling molecules such as growth factors and cytokines.

It is suggested that polymorphisms in immune-related genes may determine an individual's response to inflammation, and hence risk of cancer, at certain sites (Savage *et al.* 2004). Indeed, genetic variants have been shown to be associated with cancer susceptibility. For example, polymorphisms in cytokines *TNF-α*, *IL4*, *IL6*, *IL8* and *IL10* are reported to be associated with an increased risk of oral cancer (reviewed in Serefoglou *et al.* 2008). In oesophageal squamous cell carcinoma, variants in *TNF-α*, *TNF-β*, *IL23R, IL12A*, *IL12B*, *IL12Rβ1* and *HLA-G* have reported to be associated with the disease in Chinese populations (Guo *et al.* 2005; Chu *et al.* 2012; Tao *et al.* 2012; Chen *et al.* 2012.a), as well as a polymorphism in *IL6* in an Indian population (Upadhyay *et al.* 2008).

5.2 The Immunochip

The Immunochip is a custom genotyping array manufactured by Illumina which contains 196,524 variants that were chosen by groups researching 12 different immune-mediated inflammatory diseases (Trynka *et al.* 2011). These SNPs were either in regions of known associations in an attempt to fine-map the region, or had previously showed evidence of association but failed to meet the genome-wide significance threshold. The fine-mapping focused on 186 regions, with SNPs within 0.2 cM on either side of the top GWAS SNPs selected for inclusion on the Immunochip. These SNPs were selected from sequencing data obtained from individuals with European ancestry who were sequenced as part of the 1000 Genomes Project low-coverage pilot project, or from sequencing

data by groups involved in the chip design. Additionally, SNPs were included which showed evidence of association in non-immunological diseases, including Barrett's oesophagus and oesophageal adenocarcinoma, Parkinson's Disease, and reading and mathematical abilities, studied in the Wellcome Trust Case-Control Consortium 2 project (http://www.wtccc.org.uk/ccc2/wtccc2_studies.html). Due to the collaboration of researches in different fields, specifically in immune-related diseases, the demand for the Immunochip was high enabling the price per chip to be considerably lower than other genome-wide chips.

The large number of SNPs on the Immunochip enables the degree of population structure of sample collections to be assessed. This had not yet been investigated in the South African oesophageal cancer cases and controls, and hence, the Immunochip provides a good opportunity to do so. This, together with evidence of the involvement of inflammation in cancer development and the low-cost of the Immunochip, provided the justification for genotyping these samples on this platform.

5.3 Genotyping of South African oesophageal squamous cell carcinoma cases and controls using the Immunochip

To explore the role of immune-related genes in susceptibility to oesophageal squamous cell carcinoma in the South African Black population, 300 cases and 300 controls were genotyped on the Immunochip. This was the first large-scale genotyping study to be carried out in this population and, hence, was the first substantial investigation of population structure. A total of 300 cases and 300 controls from the South African Black population were genotyped. In addition, 50 cases and 50 controls from the South African Mixed Ancestry population were genotyped to examine population structure in this sample collection.

5.3.1 Immunochip data quality control

Several quality control (QC) steps were performed to ensure the data was of the highest quality before association tests were carried out. An initial QC step was first performed using GenomeStudio Data Analysis Software (see Methods). This removed six samples which had a genotyping call rate considerably lower than all other samples (<87% compared to >97%, respectively). Normalized intensities were then exported for the remaining 694 samples into optiCall (http://www.sanger.ac.uk/resources/software/opticall/), where the genotypes were re-called as the software is better for rare variant calling. QC steps were performed using the genotypes assigned using optiCall.

These QC measures were as follows: 1) Principal component analysis (PCA) for removal of population outliers and to reassign samples which clustered with a different population; 2) PCA to remove duplicated and highly related samples; 3) Determination of sample call rates (for each sample, this was the percentage of polymorphic variants that were assigned a genotype), and removal of samples with <95% call rate; 4) Determination of SNP call rates (for each variant, this was the percentage of samples that were assigned a genotype), and removal of those SNPs with <95% call rate.

5.3.1.1 Population stratification

Principal component analysis (PCA) was used to examine population stratification (see Methods for details). PCA infers continuous axes of genetic variation, with each axis (or eigenvector) in turn accounting for as much variation as possible (Price *et al.* 2006). As such, populations with different geographical and ethnic origins separate into clusters, enabling outliers to be identified. A total of 27,132 independent SNPs were used in the analysis. The initial PCA plot is shown in Figure 5.1.

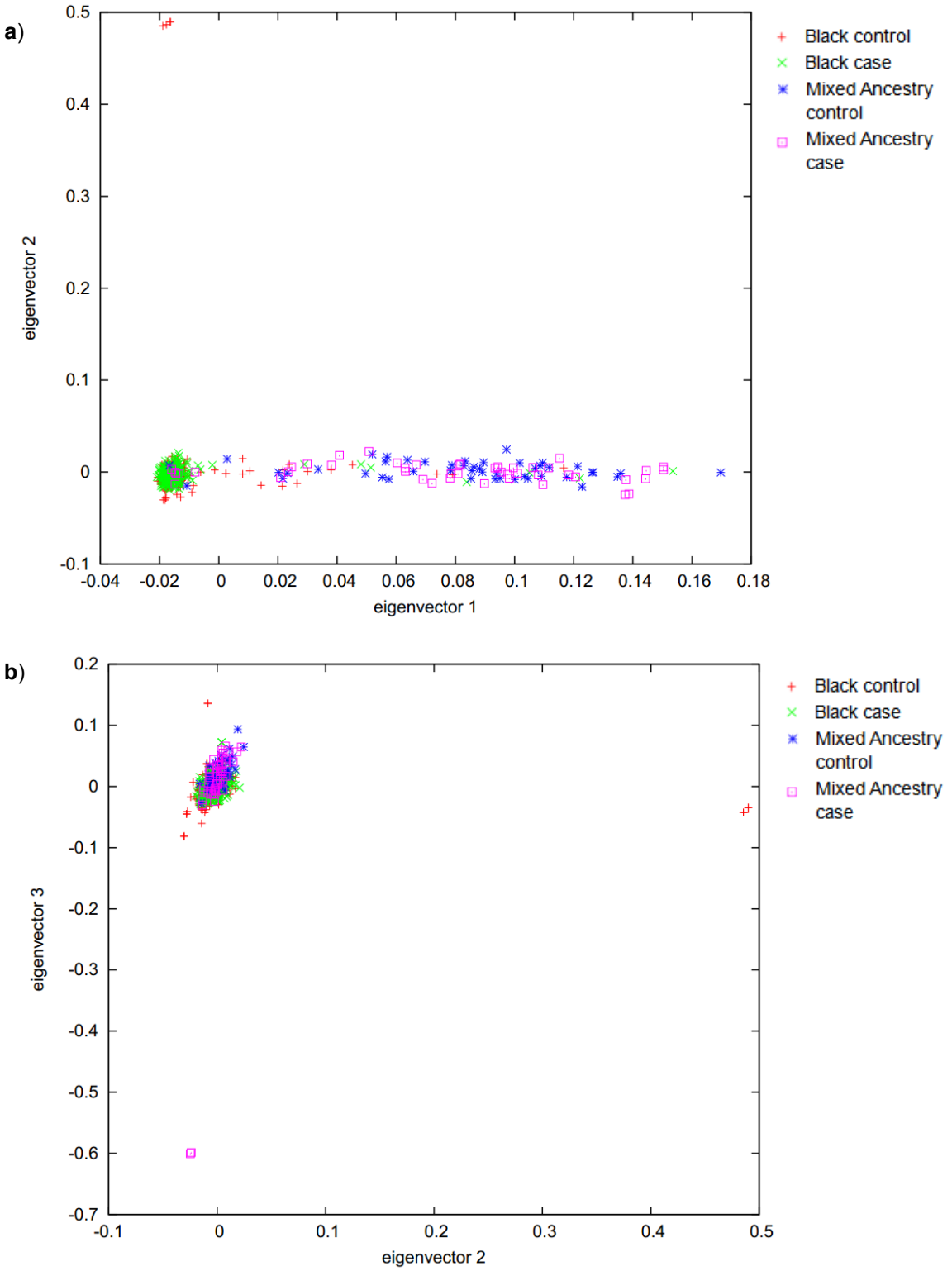


Figure 5.1: Immunochip PCA plots
Eigenvector 1 vs. eigenvector 2 (**a**), and eigenvector 2 vs. eigenvector 3 (**b**)

Figure 5.1(a) separates the South African Black and Mixed Ancestry populations along the eigenvector 1 axis, with the Black population tightly clustered and the Mixed Ancestry population showing varying Eigenvector 1 values. The majority of samples were clustered together in the Eigenvector 2 vs. 3 plot (Figure 5.1(b)), which shows that the majority of the variation between samples has already been explained by eigenvector 1 (Figure 5.1(a)). Outliers were present in both plots which were defined by thresholds of eigenvector 3 <-0.5 or eigenvector 2 >0.3, based on the eigenvector 2 vs. eigenvector 3 PCA plot (Figure 5.1(b)). Outliers may indicate a distinct population or duplicated/related samples. The level of sample relatedness for the outliers was determined (see p.158), which showed that highly related samples were present. Only one of the related samples was kept in the analysis, and the PCA was repeated, as shown in Figure 5.2. Again, outliers were present, using a threshold of eigenvector 2 values of >0.07 or <-0.06 based on the eigenvector 1 vs. eigenvector 2 PCA plot (Figure 5.2(a)). One of each highly related samples were removed.

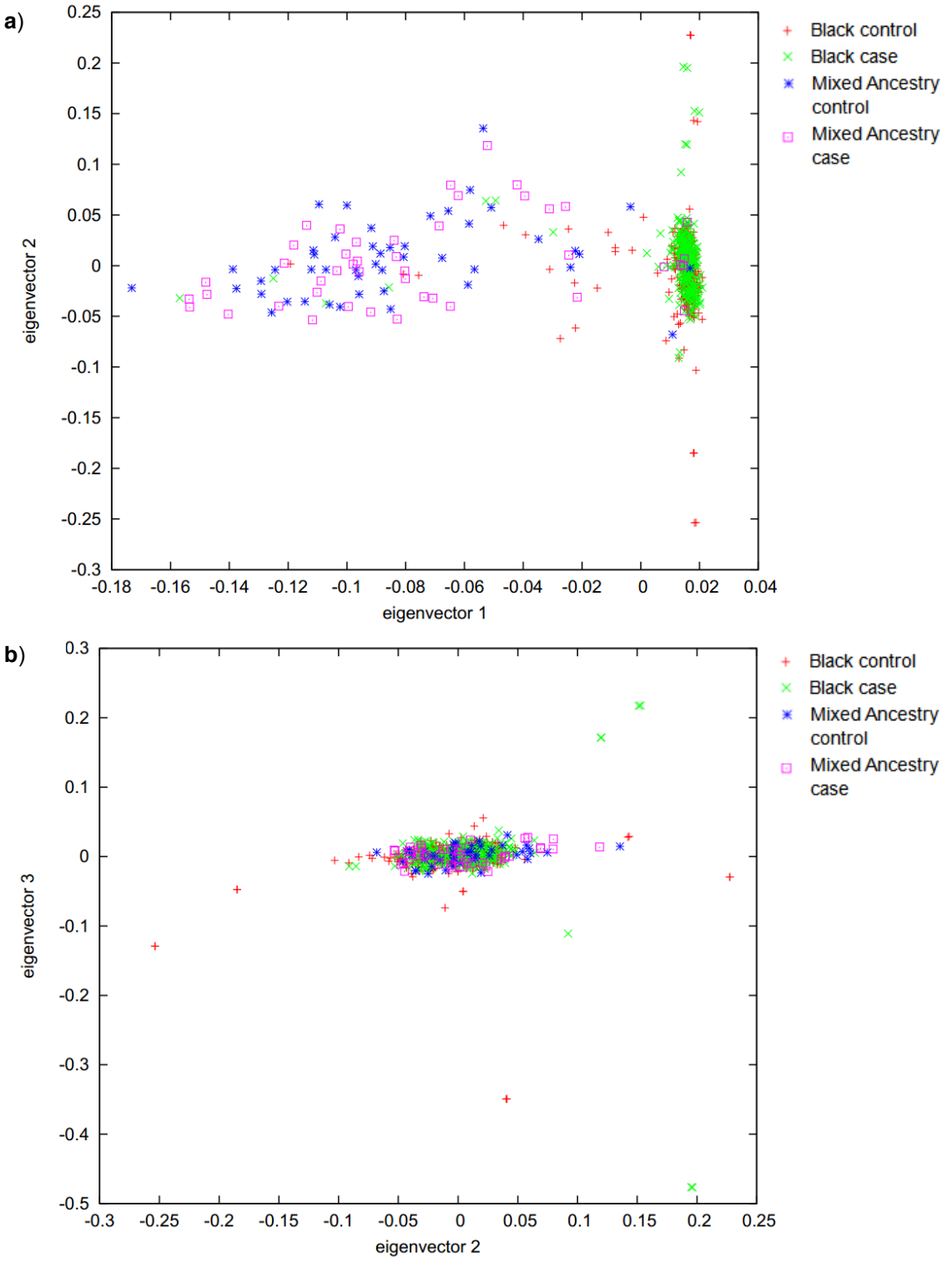


Figure 5.2: Immunochip PCA plots with outliers removed
Eigenvector 1 vs. eigenvector 2 (**a**), and eigenvector 2 vs. eigenvector 3 (**b**).

It was evident that some samples did not cluster with the population group that was self-declared by the individuals, with several Black individuals clustered with the Mixed Ancestry population, and vice versa. In total, 30 individuals clustered with a different population group using eigenvector 1=0 as a threshold (based on values in Figure 5.2(a)). For the self-declared Black population, 22 individuals clustered with the Mixed Ancestry population (7 cases and 15 controls), and 8 individuals from the Mixed Ancestry population clustered with the Black population (6 cases and 2 controls). These samples were kept in the analysis but reassigned to the population that was best described by the PCA analysis.

The South African samples were also analyzed together with several HapMap populations in a PCA to determine how they cluster with known populations (Figure 5.3). The HapMap populations were: CEU (Utah residents with Northern and Western European ancestry from the CEPH collection), ASW (African ancestry in Southwest USA), JPT (Japanese in Tokyo, Japan) and YRI (Yoruba in Ibadan, Nigeria). The South African Black population clustered tightly near the YRI population, whereas the Mixed Ancestry population showed more genetic heterogeneity than any other population in the analysis.

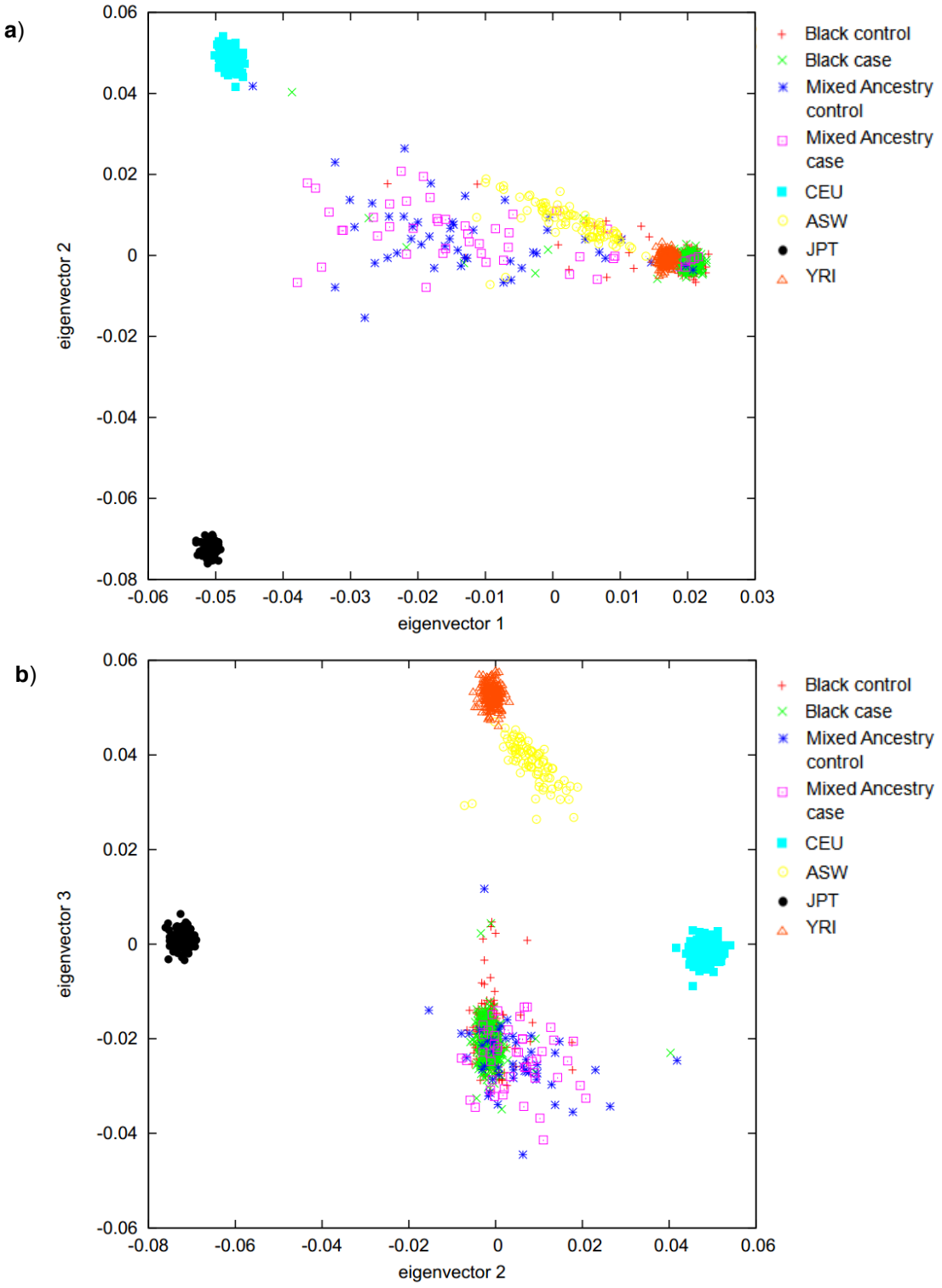


Figure 5.3: Immunochip PCA plot of the South African samples together with HapMap populations
Eigenvector 1 vs. eigenvector 2 (**a**), and eigenvector 2 vs. eigenvector 3 (**b**).
HapMap populations: CEU = Utah residents with Northern and Western European ancestry from the CEPH collection; ASW = African ancestry in Southwest USA; JPT = Japanese in Tokyo, Japan; YRI =Yoruba in Ibadan, Nigeria.

5.3.1.2 Sample genotype call rates

Sample genotype call rates (the percentage of SNPs successfully assigned a genotype for each individual) are shown in Figure 5.4 for the Black and Mixed Ancestry populations. The threshold for acceptable call rates was set at 95% for both populations. As a result of this, 30 samples were removed from the Black population, and 5 samples from the Mixed Ancestry population.

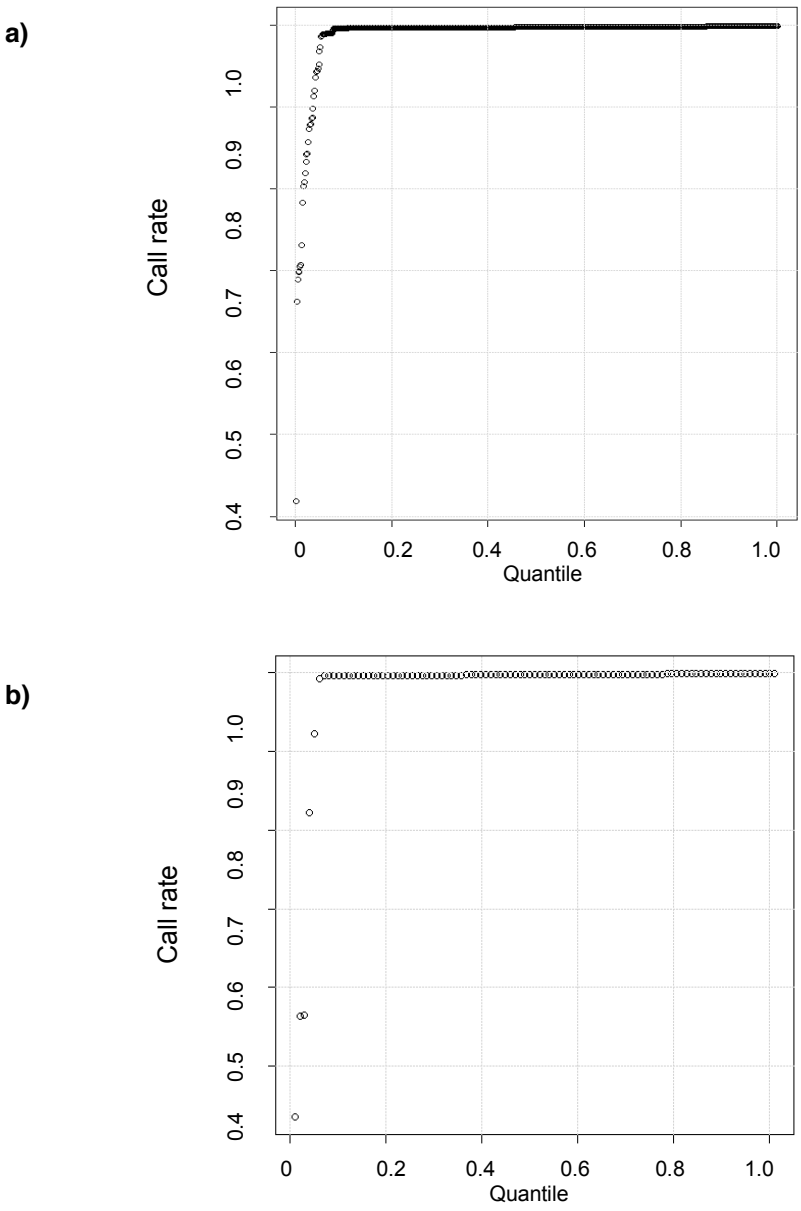


Figure 5.4: Immunochip sample genotyping call rates
Cumulative distributions for the South African Black (**a**) and Mixed Ancestry populations (**b**).

5.3.1.3 Removal of duplicated/related samples

To test if duplicated or highly related samples were present, relatedness was estimated by pairwise identity-by-descent (IBD) probabilities in PLINK using the Pi-hat score. Pi-hat is an estimate of the proportion of the genomic variation that is shared, with identical samples (e.g. monozygotic twins or duplicated samples) having a Pi-hat score of 1, first-degree relatives a Pi-hat of 0.5, and second-degree relatives 0.25.

In total, 22 pairs of individuals (21 from Black population and 1 from Mixed Ancestry population) showed a high degree of relatedness (Pi-Hat>0.5), see Table 5.1. Included in this, are four individuals who have a high degree of relatedness with more than one individual. Only one sample from each set of highly related individuals (Pi-Hat>0.5) was retained in the analysis, with the others being removed based on the sample with the lowest SNP call rates. In total, 19 individuals (18 from Black population, 1 from Mixed Ancestry population) were removed.

Table 5.1: Highly related samples from the South African populations
Highly related samples are those with Pi-hat>0.5.

Sample 1	Sample 2	Pi_Hat	Sample 1 – missing genotypes	Sample 2 – missing genotypes	Sample to remove
BN0468	BN0704	1	0.0020	0.0023	BN0704
BN0535	BN0536	0.5975	0.0035	0.0024	BN0535
BN0535	BN0770	0.6015	0.0035	0.0018	BN0535
BN0535	BN0931	1	0.0035	0.0036	BN0931
BN0536	BN0770	1	0.0024	0.0018	BN0536
BN0536	BN0931	0.589	0.0024	0.0036	BN0931
BN0580	BN0738	1	0.0037	0.0038	BN0738
BN0709	BN0810	0.6017	0.0025	0.0021	BN0709
BN0766	BN0787	0.5753	0.0028	0.0022	BN0766
BN0770	BN0931	0.5838	0.0018	0.0036	BN0931
BN0778	BN0839	0.5781	0.0020	0.0019	BN0778
BN0789	BN0863	1	0.0021	0.0019	BN0789
BN0814	BN0823	1	0.0019	0.0024	BN0823
BN0820	BN0834	0.5934	0.0024	0.0020	BN0820
BN0871	BN0875	0.5684	0.0021	0.0025	BN0875
BN0886	BN0967	1	0.0039	0.0044	BN0967
BN0913	BN0964	1	0.00174	0.002551	BN0964
BN0932	BN0937	0.6061	0.002464	0.001833	BN0932
P1394	P1397	1	0.002408	0.002125	P1394
P520	P550	1	0.002105	0.001807	P520
P549	P562	1	0.002274	0.002002	P549
P135	P195	1	0.00403	0.004585	P195

5.3.1.4 Summary of sample selection

The number of samples that were removed or added to each population is shown in Table 5.2 and Table 5.3. In the Black population, 600 samples were genotyped (300 of each cases and controls), with 535 (278 cases and 257 controls) remaining after QC. The Mixed Ancestry population began with 100 genotyped samples (50 of each cases and controls), with 106 (48 cases and 60 controls) remaining after QC.

Table 5.2: Summary of sample selection in the South African Black population

	Cases		Controls	
	Removed/ added (-/+)	Samples remaining	Removed/ added (-/+)	Samples remaining
Samples genotyped:	-	300	-	300
Removed in GenomeStudio:	-5	295	0	300
Black samples clustering with Mixed Ancestry population:	-7	288	-15	285
Mixed Ancestry samples clustering with Black population:	+6	294	+2	287
Related or duplicated samples:	-3	291	-15	272
<95% call rate:	-13	278	-15	257

Table 5.3: Summary of sample selection in the South African Mixed Ancestry population

	Cases		Controls	
	Removed/ added (-/+)	Samples remaining	Removed/ added (-/+)	Samples remaining
Samples genotyped:	-	50	-	50
Removed in GenomeStudio:	-1	49	0	50
Mixed Ancestry samples clustering with Black population:	-6	43	-2	48
Black samples clustering with Mixed Ancestry population:	+7	50	+15	63
Related or duplicated samples:	0	49	-1	62
<95% call rate:	-3	46	-2	60

5.3.1.5 SNP removal

The SNP call rates (the percentage of samples successfully assigned a genotype for each SNP) for the South African Black population is shown in Figure 5.5. The call rate threshold for inclusion in the analysis was set at 95%, and 2,559 SNPs failed to meet this threshold. A further 2,656 SNPs were removed from analysis in this population due to failure to meet the HWE threshold of $P<1x10^{-6}$, and an additional 47,301 variants were monomorphic. Therefore, a total of 139,793 SNPs were available for the case-control analysis in the Black population. A case-control association analysis was not performed in the Mixed Ancestry population, so SNP call rates are not shown.

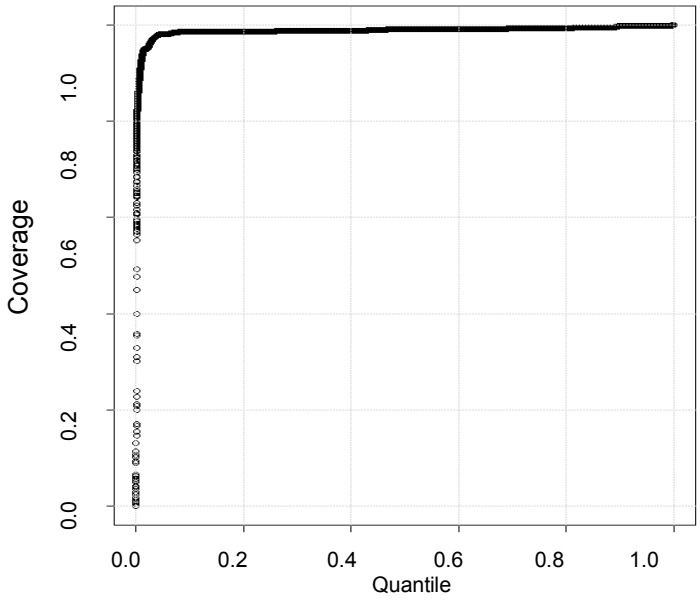


Figure 5.5: Immunochip SNP call rates
Cumulative distribution of SNP call rates in the South African Black population

5.3.1.6 Plate effects

Cases and controls were genotyped mainly on separate plates (see Table 5.4). Therefore, any differences in allele frequencies observed between cases and controls may be due to plate effects and represent false positives.

Table 5.4: Number of samples on each Immunochip genotyping plate

Plate number	Cases	Controls
5	64	25
6	0	36
7	0	83
8	0	79
9	47	32
10	89	0
11	72	0
12	2	6

To determine if plate effects were present, minor allele frequencies for SNPs with P<0.05, were compared for each plate individually against all other plates using Q-Q plots. An example of this is shown in Figure 5.6, showing the minor

allele frequencies of plate 5 against that in all the other plates (plates 6-12). No plate effects were observed, with the minor allele frequencies for each plate not deviating from that of all the other plates.

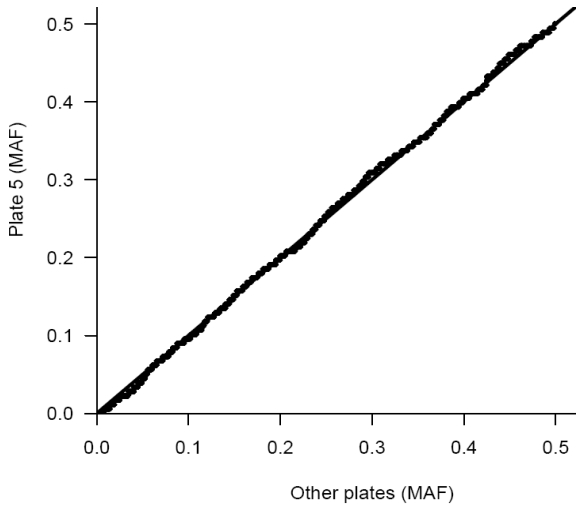


Figure 5.6: Comparison of minor allele frequencies between genotyping plates
As cases and controls were mainly genotyped on different plates, the minor allele frequencies (MAF) of SNPs with P<0.05 were plotted for each plate (plate 5 shown above) against all other plates. Plates showing different minor allele frequencies may indicate genotyping errors.

5.3.2 Case-control genetic association analysis

A case-control analysis was performed for the 139,793 SNPs in 278 cases and 257 controls from the South African Black population. After testing each SNP for association with OSCC, 221 SNPs showed at least nominal evidence of association ($P<1 \times 10^{-3}$). Genotype plots were checked for these SNPs to ensure genotypes clustered into 2 or 3 distinct groups. Based on this, the genotypes of 74 variants were found to be incorrectly clustered and, hence, were removed at this point.

A Q-Q plot showing –log(p-value) for observed results vs. expected values for all SNPs is shown in Figure 5.7. Points show a good fit for the expected line across a wide range of p-values and there is no evidence of inflation ($\lambda = 1.000$), indicating that there was no problem with population stratification in the study. At

high values of –log(p-value), several SNPs deviate from the expected P-values under the null hypothesis of no association, suggesting that there are variants associated with oesophageal cancer.

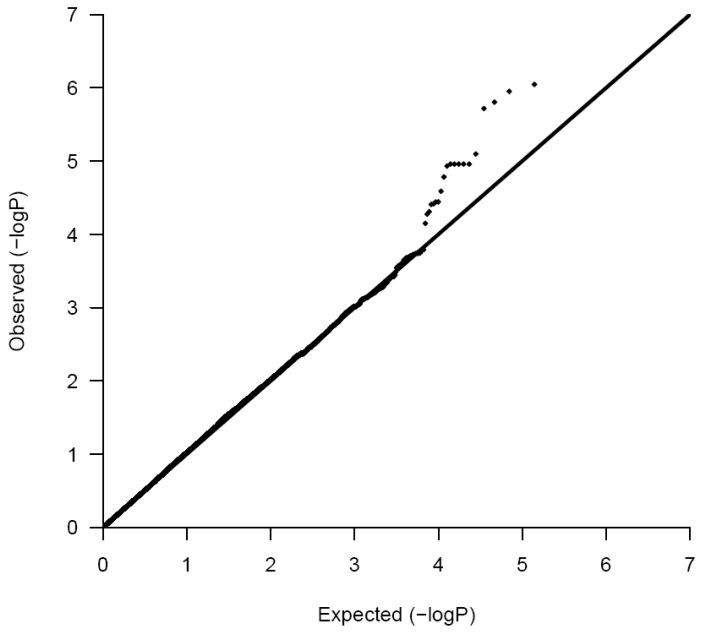


Figure 5.7: Q-Q plot of Immunochip OSCC association results
Q-Q plot showing -log of observed vs. expected p-values for all SNPs genotyped on the Immunochip (genomic control λ = 1.00).

The results of the case-control analysis for SNPs with $P_{Immunochip}<1 \times 10^{-4}$ (20 SNPs) are shown in Table 5.5. Results for SNPs with $P_{Immunochip}<1 \times 10^{-3}$ (147 SNPs) are shown in the Appendix, Table A.8.

Table 5.5: Immunochip case-control association results
SNPs with P<1 x 10^{-4}.

SNP ID	Chr	Position (b37)	Major / minor allele	MAF: Cases / controls	P-value	OR (95% CI)
rs9887787	**1**	**92222143**	**G / A**	**0.0596 / 0.1516**	**8.86 x 10^{-7}**	**0.35 (0.23-0.54)**
rs10493860	**1**	**92212703**	**G / A**	**0.0612 / 0.1529**	**1.05 x 10^{-6}**	**0.36 (0.24-0.55)**
rs2810893	**1**	**92144970**	**G / A**	**0.2392 / 0.3765**	**1.16 x 10^{-6}**	**0.52 (0.40-0.68)**
rs2182833	1	55500429	A / G	0.4245 / 0.2843	1.85 x 10^{-6}	1.86 (1.44-2.40)
rs11165441	1	92224347	G / A	0.0594 / 0.1431	4.96 x 10^{-6}	0.38 (0.25-0.58)
rs36590	22	30328070	G / A	0.0216 / 0.0807	9.51 x 10^{-6}	0.25 (0.13-0.48)
rs36596	22	30335269	G / A	0.0216 / 0.0807	9.51 x 10^{-6}	0.25 (0.13-0.48)
rs36600	22	30337586	G / A	0.0216 / 0.0807	9.51 x 10^{-6}	0.25 (0.13-0.48)
rs5763634	22	30350532	G / A	0.0216 / 0.0807	9.51 x 10^{-6}	0.25 (0.13-0.48)
rs4239932	22	30368384	C / A	0.0216 / 0.0807	9.51 x 10^{-6}	0.25 (0.13-0.48)
rs5763674	22	30386358	A / G	0.0216 / 0.0807	9.51 x 10^{-6}	0.25 (0.13-0.48)
rs5752993	22	30387160	A / G	0.0217 / 0.0807	1.01 x 10^{-5}	0.25 (0.13-0.49)
rs13147507	4	115334709	A / G	0.2464 / 0.1431	2.29 x 10^{-5}	1.96 (1.43-2.68)
rs13390918	2	199564895	G / A	0.4802 / 0.3529	2.59 x 10^{-5}	1.69 (1.32-2.17)
rs1400978	2	199565298	G / A	0.5090 / 0.3824	3.28 x 10^{-5}	1.68 (1.31-2.14)
rs12052337	2	181045431	A / G	0.1817 / 0.2882	3.89 x 10^{-5}	0.55 (0.41-0.73)
rs1001434	19	30205448	G / A	0.3435 / 0.4667	4.21 x 10^{-5}	0.60 (0.47-0.77)
rs228125	14	81338068	G / A	0.1637 / 0.0824	5.99 x 10^{-5}	2.18 (1.48-3.21)
rs1547354	21	26946709	G / A	0.0162 / 0.0635	6.64 x 10^{-5}	0.24 (0.11-0.51)
rs7714035	5	102644627	A / T	0.5036 / 0.3824	6.94 x 10^{-5}	1.64 (1.28-2.09)

Using the Bonferroni correction, a significance threshold of P<1.84 x 10^{-6} was applied to account for the multiple testing of 27,132 independent SNPs (those with r^2<0.2 and MAF >0.05). Only three SNPs were significantly associated with OSCC (as shown in bold in Table 5.5), with an additional SNP having a borderline p-value ($P_{Immunochip}$<1.85 x 10^{-6}). These variants are all located on chromosome 1, with the top three SNPs and one other within the *TGFBR3* gene. Additional clusters of associated SNPs were located on chromosome 2 (3 SNPs), and chromosome 22 (7 SNPs). The chromosomal region for each SNP with P<1 x 10^{-4} was visualised using association plots on the SNAP website which plots the association p-values together with LD between the SNP of choice and other variants. The association plots for chromosomal regions that contain multiple SNPs with P<1 x 10^{-4} are shown for rs9887787 (chr 1),

rs13390918 (chr 2) and rs5752993 (chr 22) in Figure 5.8, Figure 5.9 and Figure 5.10, respectively. The regional association plots for all SNPs with P<1 x 10^{-4} are included in the Appendix, Figure A.1.

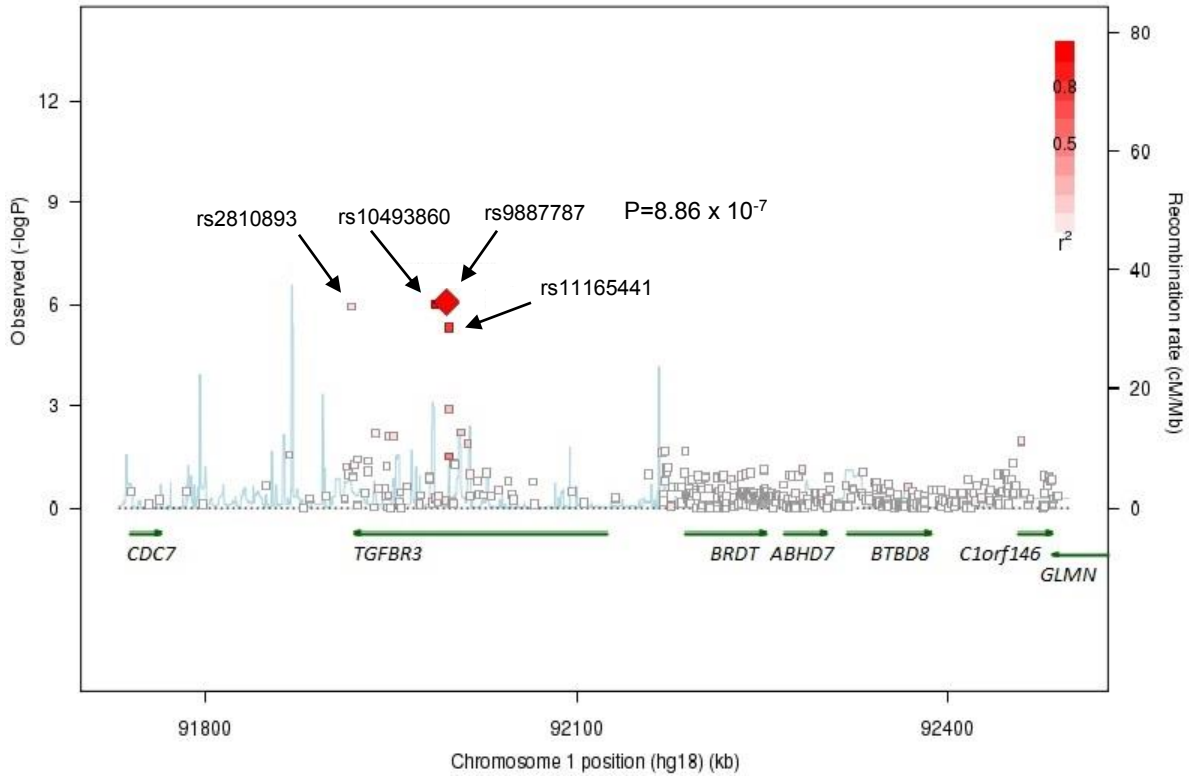


Figure 5.8: Association plot of chromosome 1 *TGFBR3* region
Shown in the diagram are the observed p-values (-logP), LD (r^2) between rs9887787 and other variants (with the colour of the square indicating the level of LD), recombination hotspots (in light blue), and known annotated genes (green).

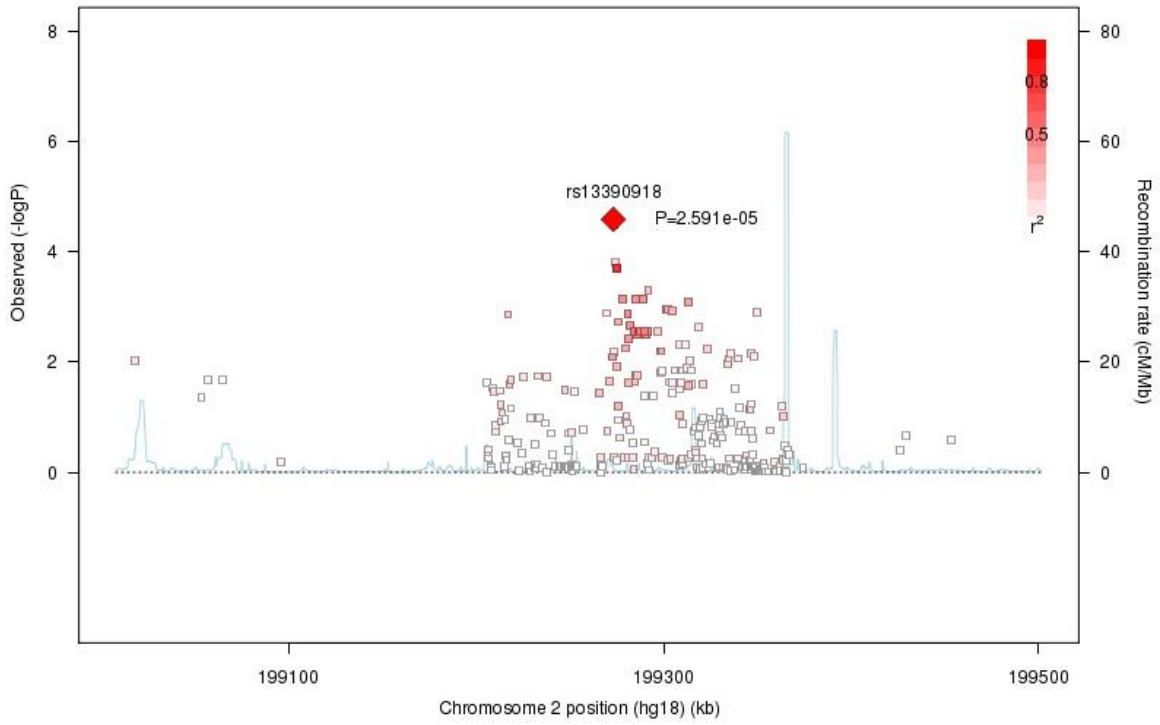


Figure 5.9: Association plot of chromosome 2 rs13390918 region
The variant rs1400978 is directly beneath the diamond for rs13390918

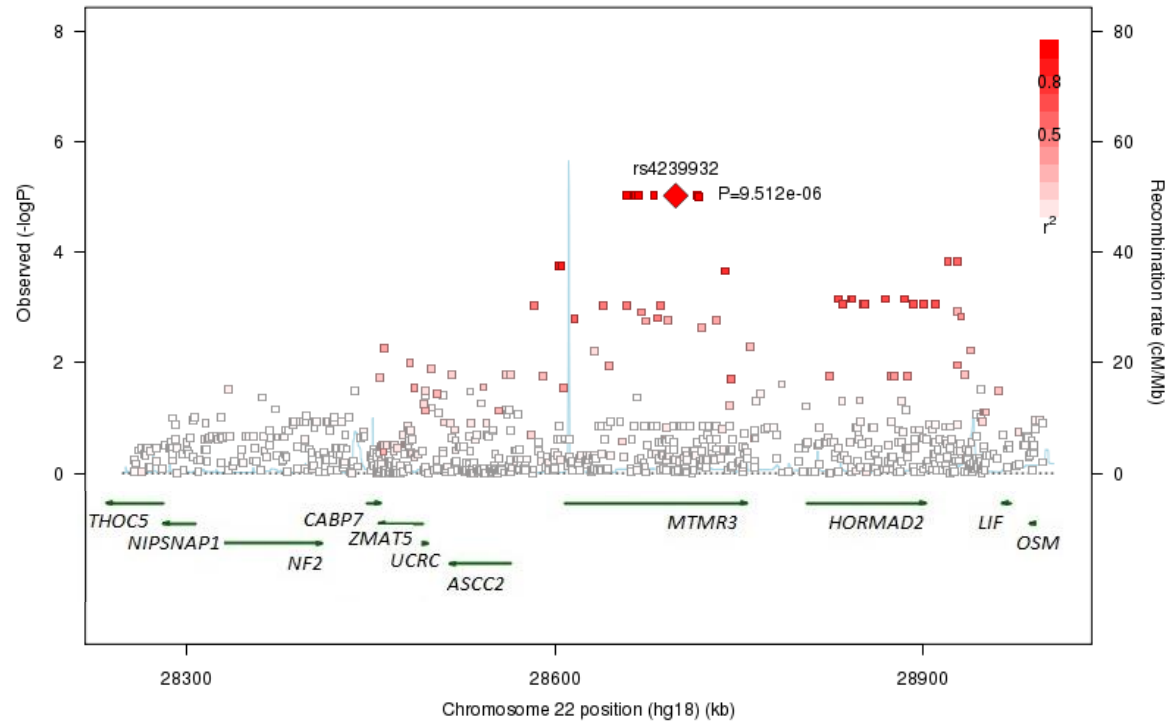


Figure 5.10: Association plot of chromosome 22 *MTMR3* rs4239932 region
This region contains 7 variants associated with OSCC that are in complete LD

The plots above all show that the association at the key SNP is supported by other SNPs in the region which have p-values of similar magnitude, as expected by the LD structure within the region, and indicating that the results are not due to genotyping errors. However, some SNPs with $P_{Immunochip}<1 \times 10^{-4}$ are in regions with few genotyped SNPs, for example rs12052337 ($P_{Immunochip}=3.89 \times 10^{-5}$) on chromosome 2, which is shown in Figure 5.11.

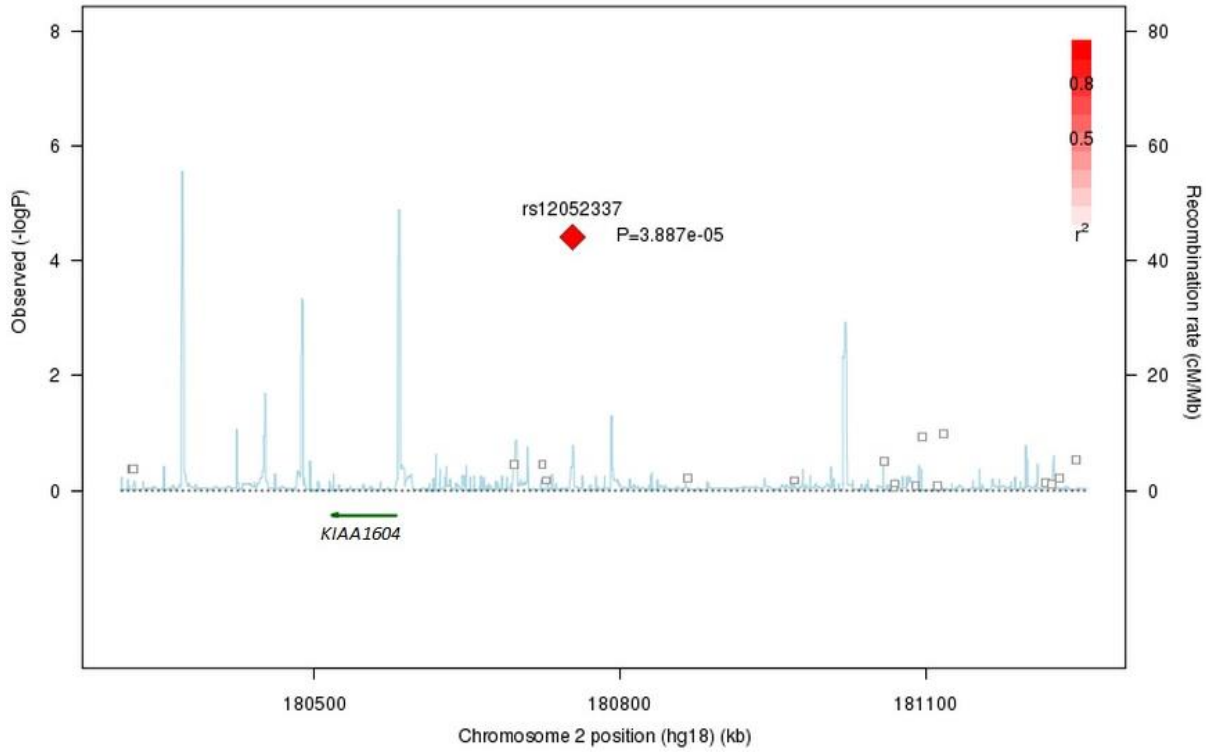


Figure 5.11: Association plot of chromosome 2 rs12052337 region

LD (r^2) was calculated between all SNPs with $P<1 \times 10^{-3}$ for the chromosomes which had multiple SNPs with $P_{Immunochip}<1 \times 10^{-4}$. LD plots are shown in Figure 5.12, Figure 5.13 and Figure 5.14 for chromosomes 1, 2 and 22, respectively.

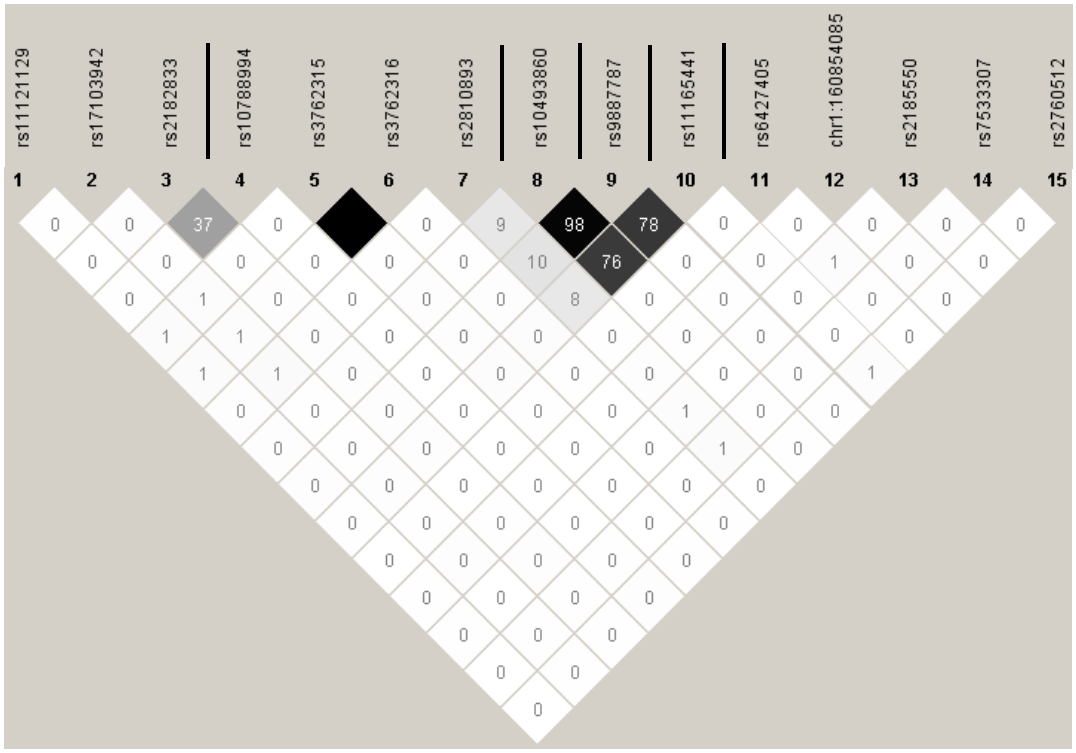


Figure 5.12: LD (r^2) between SNPs with P<0.001 on chromosome 1
SNPs with an association of P<1 x 10^{-4} are highlighted

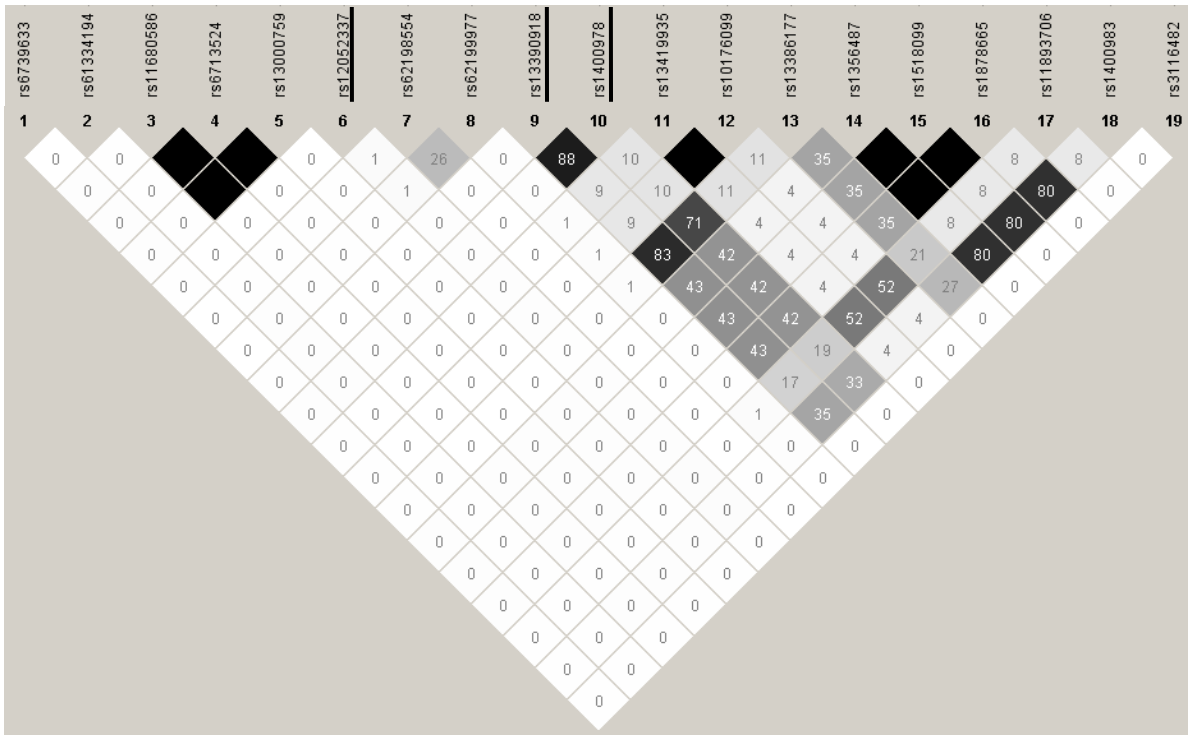


Figure 5.13: LD (r^2) between SNPs with P<0.001 on chromosome 2
SNPs with an association of P<1 x 10^{-4} are highlighted

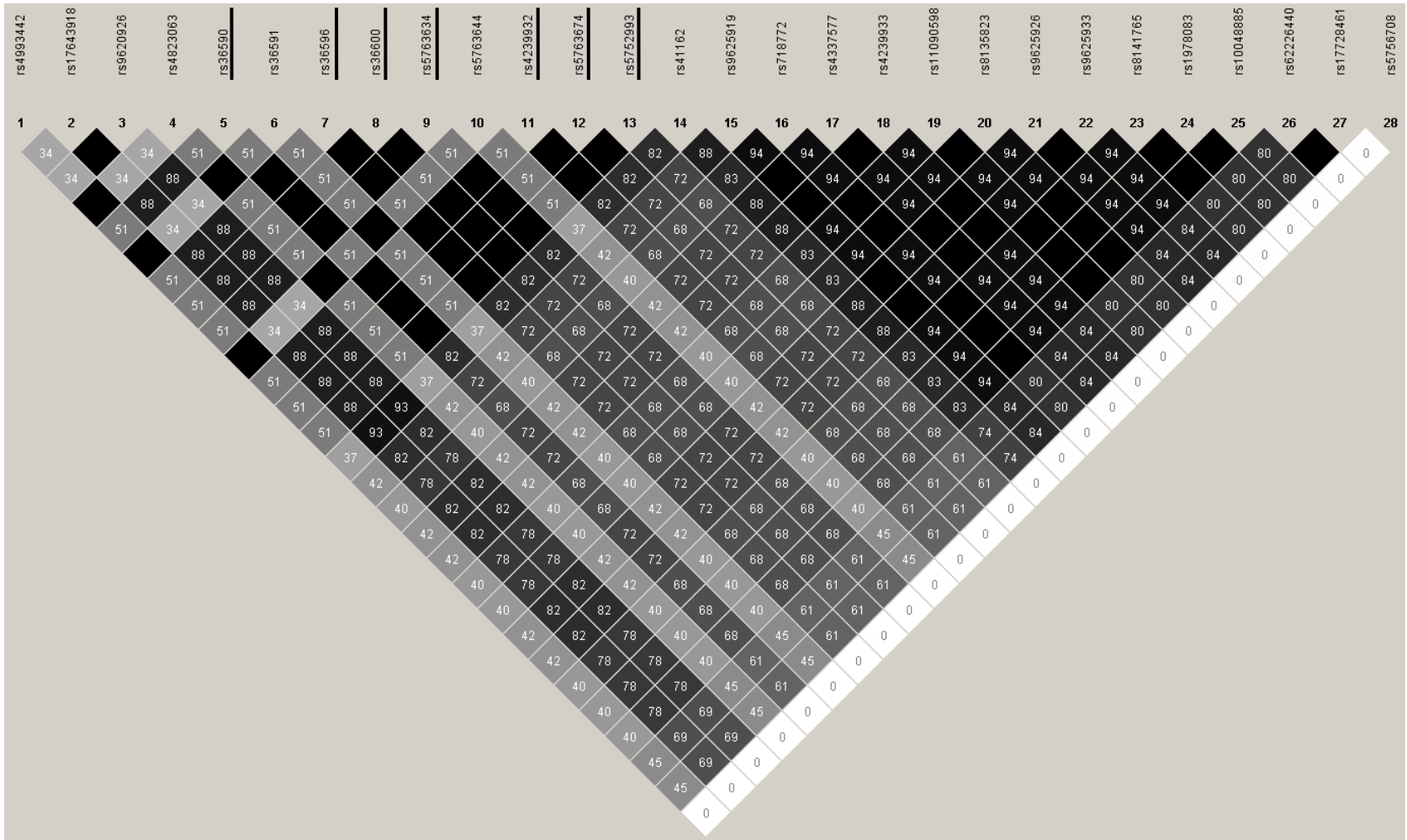


Figure 5.14: LD (r^2) between SNPs with P<0.001 on chromosome 22
SNPs with an association of P<1 x 10^{-4} are highlighted

169

On chromosome 1, three of the four SNPs at the *TGFBR3* locus with $P_{Immunochip}<1 \times 10^{-4}$ (rs9887787, rs10493860 and rs11165441) are in high LD ($r^2>0.7$). The other SNP, rs2810893, is not in LD with these (r^2 = 0.08-0.1), which suggests that it may be an independent signal. However, using logistic regression and conditioning on the top SNP rs9887787, rs2810893 had an association of $P=1.0 \times 10^{-3}$ which is less significant than the non-conditioned result ($P=1.16 \times 10^{-6}$) suggesting that it is not independent. The associations for rs10493860 and rs11165441 disappear when conditioned on rs9887787. The other SNP on chromosome 1 showing nominal evidence of disease association, rs2182833, is located 37 Mb away from the *TGFBR3* locus so is likely to be an independent association.

On chromosome 22, the majority of the 28 variants are in a high level of LD, including the 7 SNPs with $P<1 \times 10^{-4}$ which are in complete LD ($r^2=1$). These 7 SNPs are also in high LD ($r^2>0.87$) in the Yoruban population (using the 1000 Genomes dataset).

5.3.3 Extension association study for selected variants

Seven variants were chosen for genotyping in an extension study in the South African Black population. An independent replication was not possible due to insufficient sample sizes to perform a well-powered study. Therefore, all available samples (404 OSCC cases and 834 controls) were genotyped to maximize the power to detect associations. These included the 278 cases and 257 controls that were analyzed in the Immunochip assay, although the samples were genotyped again using TaqMan assays.

SNPs with an association with OSCC of $P_{Immunochip} < 1 \times 10^{-4}$ were considered for this extension study, with variants selected based on: (1) having the lowest p-values; (2) the SNAP association plots (SNPs were prioritized if the association was supported by nearby variants with p-values of similar magnitude); (3) LD

between associated SNPs (with one SNP selected from variants in high LD). The selected SNPs are shown in Table 5.6.

Table 5.6: Genotyped SNPs for Immunochip extension study

Variant	Chr	Position (build 37)	Gene	Location
rs9887787	1	92222143	*TGFBR3*	Intronic
rs2810893	1	92144970	*TGFBR3*	Downstream
rs2182833	1	55500429	*PCSK9*	Upstream
rs13390918	2	199564895	*Intergenic*	-
rs13147507	4	115334709	*Intergenic*	-
rs7714035	5	102644627	*Intergenic*	-
rs36590	22	30328070	*MTMR3*	Intronic

The TaqMan genotyping call rate was >97.8% for all SNPs. Genotypes for all SNPs were in HWE in controls (P>0.5), and as well as in cases (P>0.1) except for rs9887787 (P=0.0091). Genotypic and allelic association results are shown in Table 5.7.

Table 5.7: Genotypic and allelic association results for the Immunochip extension study in the South African Black population

Variant	Genotype or allele	Cases		Controls		OR (95% CI)	P-value
rs9887787	G/G	350	0.879	624	0.759	Ref	Ref
	G/A	43	0.108	185	0.225	0.41 (0.29 - 0.59)	1.0×10^{-6}
	A/A	5	0.013	13	0.016	0.69 (0.24 - 1.94)	0.155
	G	743	0.933	1433	0.872	Ref	Ref
	A	**53**	**0.067**	**211**	**0.128**	**0.48 (0.35 - 0.66)**	**4.0×10^{-6}**
rs2810893	G/G	234	0.584	380	0.461	Ref	Ref
	G/A	137	0.342	355	0.430	0.63 (0.49 - 0.81)	3.3×10^{-4}
	A/A	30	0.075	90	0.109	0.54 (0.35 - 0.84)	6.2×10^{-3}
	G	605	0.754	1115	0.676	Ref	Ref
	A	**197**	**0.246**	**535**	**0.324**	**0.68 (0.56 - 0.82)**	**6.6×10^{-5}**
rs2182833	A/A	135	0.338	370	0.447	Ref	Ref
	A/G	200	0.500	360	0.435	1.52 (1.17 - 1.98)	1.6×10^{-3}
	G/G	65	0.163	97	0.117	1.84 (1.27 - 2.66)	1.2×10^{-3}
	A	470	0.588	1100	0.665	Ref	Ref
	G	**330**	**0.413**	**554**	**0.335**	**1.39 (1.17 - 1.66)**	**1.8×10^{-4}**
rs13390918	G/G	120	0.304	306	0.373	Ref	Ref
	A/G	193	0.489	391	0.476	1.26 (0.96 - 1.65)	0.098
	A/A	82	0.208	124	0.151	1.69 (1.19 - 2.39)	3.3×10^{-3}
	G	433	0.548	1003	0.611	Ref	Ref
	A	**357**	**0.452**	**639**	**0.389**	**1.29 (1.09 - 1.54)**	**3.2×10^{-3}**
rs13147507	A/A	235	0.595	548	0.672	Ref	Ref
	A/G	132	0.334	241	0.296	1.28 (0.98 - 1.66)	0.066
	G/G	28	0.071	26	0.032	2.51 (1.44 - 4.38)	8.3×10^{-4}
	A	602	0.762	1337	0.820	Ref	Ref
	G	**188**	**0.238**	**293**	**0.180**	**1.43 (1.16 - 1.75)**	**7.7×10^{-4}**
rs7714035	A/A	110	0.277	304	0.370	Ref	Ref
	A/T	191	0.481	391	0.476	1.35 (1.02 - 1.78)	0.034
	T/T	96	0.242	126	0.153	2.11 (1.49 - 2.97)	1.8×10^{-5}
	A	411	0.518	999	0.608	Ref	Ref
	T	**383**	**0.482**	**643**	**0.392**	**1.45 (1.22 - 1.72)**	**2.1×10^{-5}**
rs36590	G/G	380	0.948	729	0.880	Ref	Ref
	G/A	21	0.052	96	0.116	0.42 (0.26 - 0.68)	3.5×10^{-4}
	A/A	0	0.000	3	0.004	-	-
	G	781	0.974	1554	0.938	Ref	Ref
	A	**21**	**0.026**	**102**	**0.062**	**0.41 (0.25 - 0.66)**	**1.6×10^{-4}**

A summary of the allelic association results together with the corresponding data from the Immunochip for comparison is shown in Table 5.8. The disease association was less significant in the extension study as compared to the Immunochip data for six of the seven SNPs. One SNP, rs7714035, had strengthened evidence of association from $P_{Immunochip} = 6.94 \times 10^{-5}$ to $P_{Extension} = 2.1 \times 10^{-5}$. The SNP with the strongest evidence of association is rs9887787, with $P_{Extension} = 4.0 \times 10^{-6}$. However, if the Bonferroni correction is applied for all independent SNPs tested on the Immunochip ($P < 1.84 \times 10^{-6}$), none of the variants were significantly associated with OSCC. The less significant results in the extension study are mainly a result of changes in the allele frequencies in the expanded panel of controls. Nonetheless, the direction of the effects reflected in the odds ratios, remain the same.

Table 5.8: Summary of allelic association results for the Immunochip extension study and Immunochip data

Variant	Major allele	Minor allele	Extension study			Immunochip study		
			MAF: cases/controls	OR (95% CI)	P-value	MAF: cases/controls	OR (95% CI)	P-value
rs9887787	G	A	0.067 / 0.128	0.48 (0.35 - 0.66)	4.0×10^{-6}	0.060 / 0.152	0.35 (0.23 - 0.54)	8.86×10^{-7}
rs2810893	G	A	0.246 / 0.324	0.68 (0.56 - 0.82)	6.6×10^{-5}	0.239 / 0.377	0.52 (0.40 - 0.68)	1.16×10^{-6}
rs2182833	A	G	0.413 / 0.335	1.39 (1.17 - 1.66)	1.8×10^{-4}	0.425 / 0.284	1.86 (1.44 - 2.40)	1.85×10^{-6}
rs13390918	G	A	0.452 / 0.389	1.29 (1.09 - 1.54)	3.2×10^{-3}	0.480 / 0.353	1.69 (1.32 - 2.17)	2.59×10^{-5}
rs13147507	A	G	0.238 / 0.180	1.43 (1.16 - 1.75)	7.7×10^{-4}	0.246 / 0.143	1.96 (1.43-2.68)	2.29×10^{-5}
rs7714035	A	T	0.482 / 0.392	1.45 (1.22 - 1.72)	2.1×10^{-5}	0.504 / 0.382	1.64 (1.28 - 2.09)	6.94×10^{-5}
rs36590	G	A	0.026 / 0.062	0.41 (0.25 - 0.66)	1.6×10^{-4}	0.022 / 0.081	0.25 (0.13 - 0.48)	9.51×10^{-5}

5.3.4 Comparison of TaqMan and Immunochip genotypes

Genotyping the same samples and variants on two independent platforms allows comparison between the methods. Overall, 3745 genotypes were tested (535 samples for each of the seven variants). Of these, 37 genotypes could not be called in the TaqMan SNP assays, and 3 were not called on the Immunochip (0.99% and 0.08%, respectively). Only one genotype that was called on both platforms produced a conflicting result; sample P332 for rs13390918 had an AA genotype using Immunochip but GG genotype using the TaqMan assay.

5.3.5 Gene-environment interactions

Cases and controls from the extension study were stratified by smoking status and alcohol consumption to test for gene-environment interactions for all seven SNPs genotyped in the extension study. Four tests were carried out for each risk factor. For example, for gene-alcohol interactions the tests were: drinkers case-control, non-drinkers case-control, case-only (drinker vs. non-drinker), and a gene-alcohol interaction test using logistic regression controlling for age, sex, smoking, alcohol, the genetic effect and the gene-alcohol interaction effect. Results are shown in Table 5.9 and Table 5.10. The Bonferroni correction used for the original case-control analysis was used as the significance threshold ($P<1.84 \times 10^{-6}$), except for the gene-environment interaction test, which used a threshold of $P<0.0071$ (0.05/7).

Table 5.9: Gene-alcohol interaction tests for SNPs genotyped in the Immunochip extension study
Stratified analysis of alcohol consumption, a case-only analysis, and a gene-environment (G x E) interaction test using logistic regression. Cases: drinkers = 252, non-drinkers = 149; Controls: drinkers = 444, non-drinkers = 386.

		Case-control: Drinkers only				Case-control: Non-drinkers only			
Variant	Minor allele	MAF: Cases	MAF: Controls	OR (95% CI)	P-value	MAF: Cases	MAF: Controls	OR (95% CI)	P-value
rs36590	T	0.024	0.061	0.38 (0.20 - 0.71)	0.0018	0.03	0.063	0.47 (0.23 - 0.97)	0.0365
rs2182833	G	0.41	0.312	1.53 (1.22 - 1.92)	0.0003	0.419	0.362	1.27 (0.97 - 1.67)	0.0861
rs2810893	T	0.244	0.317	0.69 (0.54 - 0.89)	0.0040	0.243	0.329	0.66 (0.48 - 0.89)	0.0065
rs7714035	T	0.478	0.389	1.44 (1.15 - 1.80)	0.0013	0.493	0.395	1.49 (1.13 - 1.95)	0.0040
rs9887787	T	0.071	0.12	0.56 (0.37 - 0.83)	0.0040	0.058	0.137	0.39 (0.23 - 0.66)	0.0003
rs13390918	T	0.467	0.386	1.40 (1.12 - 1.75)	0.0034	0.418	0.395	1.10 (0.84 - 1.45)	0.4783
rs13147507	G	0.227	0.179	1.35 (1.02 - 1.77)	0.0328	0.253	0.179	1.55 (1.13 - 2.14)	0.0070

		Case-only				G x E	
Variant	Minor allele	MAF: Drinkers	MAF: Non-drinkers	OR (95% CI)	P-value	OR (95% CI)	P-value
rs36590	T	0.024	0.03	0.78 (0.33 - 1.88)	0.5858	0.74 (0.27 - 2.03)	0.5562
rs2182833	G	0.41	0.419	0.96 (0.72 - 1.29)	0.7973	1.38 (0.94 - 2.02)	0.0989
rs2810893	T	0.244	0.243	1.00 (0.72 - 1.40)	1.000	1.04 (0.70 - 1.55)	0.8506
rs7714035	T	0.478	0.493	0.94 (0.70 - 1.26)	0.6775	1.03 (0.71 - 1.49)	0.8772
rs9887787	T	0.071	0.058	1.24 (0.68 - 2.25)	0.4851	1.20 (0.62 - 2.32)	0.5859
rs13390918	T	0.467	0.418	1.22 (0.91 - 1.63)	0.1819	1.02 (0.71 - 1.47)	0.9230
rs13147507	G	0.227	0.253	0.87 (0.62 - 1.21)	0.4082	0.63 (0.40 - 0.99)	0.0446

Table 5.10: Gene-smoking interaction tests for SNPs genotyped in the Immunochip extension study
Stratified analysis of tobacco smoking status, a case-only analysis, and a gene-environment (G x E) interaction test using logistic regression. Cases: smokers = 240, non-smokers = 163; Controls: smokers = 327, non-smokers = 496.

		Case-control: Smokers only				Case-control: Non-smokers only			
Variant	**Minor allele**	**MAF: Cases**	**MAF: Controls**	**OR (95% CI)**	**P-value**	**MAF: Cases**	**MAF: Controls**	**OR (95% CI)**	**P-value**
rs36590	T	0.025	0.064	0.38 (0.20 - 0.72)	0.0023	0.028	0.061	0.44 (0.22 - 0.89)	0.0201
rs2182833	G	0.418	0.289	1.77 (1.38 - 2.27)	7.0×10^{-6}	0.401	0.368	1.15 (0.89 - 1.49)	0.2825
rs2810893	T	0.235	0.323	0.65 (0.49 - 0.85)	0.0014	0.262	0.326	0.74 (0.56 - 0.98)	0.033
rs7714035	T	0.460	0.394	1.31 (1.03 - 1.66)	0.0283	0.513	0.388	1.66 (1.29 - 2.14)	8.4×10^{-5}
rs9887787	T	0.078	0.136	0.54 (0.36 - 0.81)	0.0026	0.050	0.121	0.38 (0.22 - 0.65)	2.6×10^{-4}
rs13390918	T	0.459	0.399	1.28 (1.02 - 1.60)	0.0300	0.441	0.375	1.31 (1.00 - 1.73)	0.0484
rs13147507	G	0.231	0.186	1.31 (0.98 - 1.76)	0.0694	0.250	0.177	1.55 (1.15 - 2.10)	0.0039

		Case-only				G x E	
Variant	**Minor allele**	**MAF: Smokers**	**MAF: non-smokers**	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**
rs36590	T	0.025	0.028	0.91 (0.38 - 2.17)	0.8235	0.99 (0.36 - 2.74)	0.9835
rs2182833	G	0.418	0.401	1.07 (0.80 - 1.43)	0.642	1.74 (1.20 - 2.53)	**0.0037**
rs2810893	T	0.235	0.262	0.87 (0.62 - 1.20)	0.3833	1.11 (0.74 - 1.66)	0.6095
rs7714035	T	0.460	0.513	0.81 (0.61 - 1.08)	0.1448	0.81 (0.56 - 1.16)	0.2442
rs9887787	T	0.078	0.050	1.63 (0.89 - 2.98)	0.1116	1.44 (0.72 - 2.88)	0.3058
rs13390918	T	0.459	0.441	1.08 (0.81 - 1.43)	0.6132	1.14 (0.79 - 1.64)	0.4827
rs13147507	G	0.231	0.25	0.90 (0.65 - 1.25)	0.5296	0.79 (0.51 - 1.22)	0.2894

No variants were significantly associated with drinking status, with the gene-alcohol interaction test and the case-only analysis showing no evidence of interaction.

One variant, rs2182833, appeared to interact with smoking status. The minor allele frequencies for cases who are smokers and non-smokers are similar (0.418 and 0.401, respectively), but the frequencies in controls for these groups differ (0.289 and 0.368, respectively). Evidence for association of this variant with OSCC was strengthened in smokers ($P = 7.0 \times 10^{-6}$; OR = 1.77), compared to the non-stratified analysis ($P = 1.8 \times 10^{-4}$; OR = 1.39), see Table 5.7. The gene-smoking interaction test for this variant gives P = 0.0037, which is significant when using the Bonferroni correction for the smoking interactions only (P<0.0071).

Additionally, an Immunochip-wide gene-environment interaction analysis (using 278 cases and 257 controls) was performed for alcohol consumption and smoking (see Table 5.11 and Table 5.12, respectively, for the 10 SNPs with the strongest evidence for interaction). No variants showed an association with OSCC which would survive the Bonferroni correction for multiple testing ($P< 1.84 \times 10^{-6}$).

Table 5.11: Immunochip-wide gene-alcohol interaction test

Chr	SNP ID	Position (b37)	OR (95% CI)	P-value
1	rs10493941	101445821	3.22 (1.74 - 5.93)	1.83×10^{-4}
1	rs12726628	101446276	3.22(1.74 - 5.93)	1.83×10^{-4}
2	rs17025022	40291538	0.33 (0.18 - 0.59)	$2.13x\ 10^{-4}$
1	rs17123572	101442211	3.17 (1.72 - 5.84)	2.21×10^{-4}
1	rs2494287	203207520	8.38 (2.51 - 27.96)	5.45×10^{-4}
1	rs1400986	207038686	2.60 (1.51 - 4.49)	6.14×10^{-4}
1	rs4908109	101453655	0.31 (0.16 - 0.61)	6.93×10^{-4}
1	rs1188734	101457021	0.31 (0.16 - 0.61)	6.93×10^{-4}
1	rs67858127	17678599	0.35 (0.19 - 0.65)	9.05×10^{-4}
1	rs7542831	101472742	2.53 (1.46 - 4.39)	9.15×10^{-4}

Table 5.12: Immunochip-wide gene-smoking interaction test

Chr	SNP ID	Position (b37)	OR (95% CI)	P-value
1	rs1052240	198634814	3.49 (1.98 - 6.16)	1.66×10^{-5}
7	rs10085839	131570695	0.25 (0.13 - 0.48)	4.26×10^{-5}
11	rs11215001	114338079	3.02 (1.73 - 5.26)	1.02×10^{-4}
11	rs7124064	114344225	3.02 (1.73 - 5.26)	1.02×10^{-4}
10	rs4934742	35559191	0.31 (0.17 - 0.57)	1.23×10^{-4}
11	rs10750052	114342631	0.34 (0.20 - 0.59)	1.24×10^{-4}
8	1kg_8_11087411*	11050001	4.71 (2.12 - 10.46)	1.41×10^{-4}
16	rs16961198	59591652	0.24 (0.11 - 0.50)	1.41×10^{-4}
1	rs7515488	1163804	3.17 (1.74 - 5.77)	1.61×10^{-4}
2	rs759844	207428868	2.69 (1.60 - 4.53)	2.00×10^{-3}

*Immunochip ID number shown as an 'rs' number was not available

5.4 Replication of OSCC GWAS associated SNPs in the South African Black population using Immunochip

To date, three independent OSCC GWAS have been completed in the Chinese population (Abnet *et al.* 2010; Wang *et al.* 2010.a; Wu *et al.* 2011.c). A meta-analysis has also been performed for two of these (Abnet *et al.* 2012) and additionally, a more thorough replication of the Wu *et al.* GWAS has been published, whereby variants with $10^{-7} < P < 10^{-4}$ were studied (Wu *et al.* 2012.a). Based on these five studies, 38 SNPs were significantly associated with OSCC, although not all were independent loci. A GWAS for Barrett's oesophagus in a European population has also identified two variants associated with the disease (Su *et al.* 2012).

The degree of association with OSCC in the South African Black population of these 40 variants was examined using the Immunochip data. In addition to the index SNPs identified in the published studies, tagging variants (or proxies) were also explored. To identify variants in high LD with the index SNPs (r^2>0.8 and within 500kb) that were present on the Immunochip, a SNAP proxy search was used (http://www.broadinstitute.org/mpg/snap/ldsearch.php). This was performed in the Chinese/Japanese (CHB/JPT) or European (CEU) HapMap 3 and 1000 Genomes populations, for OSCC studies or the Barrett's oesophagus

GWAS, respectively. A proxy search was also performed in the Yoruban (YRI) population for both the index SNPs and proxies identified in the CHB/JPT and CEU populations. This was to identify variants that were in high LD with the index SNP in the YRI population, and to determine if the proxies identified in the CHB/JPT and CEU populations also tag additional variants in the YRI population. However, for the latter, no additional proxies were identified. The 40 index SNPs, together with any proxies present on the Immunochip are shown in Table 5.13.

Table 5.13: Summary of OSCC and Barrett's oesophagus GWAS associations, and the presence of these index SNPs or proxies on the Immunochip

Study	Chr location	Gene or region	rs number	Index SNP or proxy on Immuno-chip	SNPs on Immunochip that are in high LD (r^2>0.8) with index SNP in 1000 Genomes or HapMap 3 populations	
					CHB/JPT or CEU* (r^2)	YRI (r^2)
Abnet *et al.* 2010	10q23	*PLCE1*	rs2274223	No	-	-
			rs3765524	No	-	-
			rs3781264	No	-	-
			rs11187842	No	-	-
			rs753724	No	-	-
	22q12	*CHEK2*	rs738722	No	-	-
Wang *et al.* 2010	10q23	*PLCE1*	rs12263737	No	-	-
			rs2274223	No	-	-
	20p13	*C20orf54*	rs13042395	No	-	-
Wu *et al.* 2011	10q23	*PLCE1*	rs2274223	No	-	-
	5q11	*PDE4D*	rs10052657	No	-	-
	21q22	*RUNX1*	rs2014300	No	-	-
	6p21	*near UNC5CL*	rs10484761	No	-	-
	12q24	*ACAD10*	rs11066015	No	-	-
	12q24	*near RPL6 and PTPN11*	rs11066280	No	-	-
	12q24	*C12orf51*	rs2074356	Index SNP	-	-
Abnet *et al.* 2012	2q33.1	*CASP8/ ALS2CR12/TRAK2*	rs10931936	Proxies	rs13016963 (0.96) rs9288316 (0.96)	-
			rs13016963	Index SNP, proxies	rs9288316 (1.00)	-
			rs9288318	Proxies	rs13016963 (0.83) rs9288316 (0.82)	-
			rs10201587	Proxies	rs13016963 (0.83) rs9288316 (0.82)	-
			rs7578456	No	-	-

Wu *et al.* 2012	16q12.1	*HEATR3*	rs4785204	Proxies	rs12445755 (0.96) rs7204293 (0.89)	-
			rs7206735	Proxies	rs6500291 (1.00)	-
	17q21	*HAP1*	rs6503659	No	-	-
	22q12	*XBP1*	rs2239815	Proxies	rs2097461 (0.97) rs5762788 (1.00) rs5762795 (1.00)	rs5762795 (0.83)
	3q27	*ST6GAL1*	rs2239612	No	-	-
	17p13	*SMG6*	rs17761864	No	-	-
	18p11	*PTPN2*	rs2847281	Index SNP	-	-
	22q12	*CHEK2*	rs4822983	No	-	-
			rs1033667	No	-	-
	4q23	*ADH1A*	rs1229977	No	-	-
		ADH1B	rs1042026	Proxies	rs1229984 (0.89)	-
			rs17033	No	-	-
		ADH1C	rs1614972	Index SNP	-	-
			rs1789903	No	-	-
		ADH4	rs3805322	No	-	-
		ADH7	rs17028973	No	-	-
		ADH6	rs1893883	No	-	-
	2q22	*IGFB2*	rs9288520	No	-	-
	13q33	*SLC10A2*	rs17450420	No	-	-
Su *et al.* 2012	6p21	*MHC region*	rs9257809	Index SNP, proxies	rs429479 (1.00)* rs406511 (1.00)* rs1233480 (0.95)* rs442694 (0.95)* rs1535039 (0.95)* rs2746149 (0.95)* rs2746150 (0.95)* rs1233491 (0.95)* rs2523443 (0.91)* rs404240 (0.91)* rs3749971 (0.86)* rs3117427 (0.81)* rs3117439 (0.81)* rs3117425 (0.81)*	rs406511 (0.871)
	16q24	*Near FOX1*	rs9936833	Index SNP, proxies	rs1532167 (1.00)* rs7186259 (0.86)*	rs1532167 (0.93) rs7186259 (0.84)

Only 6 variants were present on the Immunochip, with association results only available for four SNPs; rs13016963 in the region on chromosome 2q containing *CASP8*/*ALS2CR12*/*TRAK2*, rs2847281 in *PTPN2*, rs1614972 in *ADH1C* and rs9936833 near *FOX1*. The other 2 variants were monomorphic in controls and/or cases. In total, ten index SNPs had proxies that were present on the Immunochip. The association results for the index SNPs and proxies are shown in Table 5.14. None of the variants were associated with

OSCC in the South African Black population (P > 0.0.5), with several SNPs being monomorphic in this population.

Table 5.14: Immunochip OSCC association results for the South African Black population for SNPs previously associated with OSCC and Barrett's oesophagus in GWAS

Index SNP	Tagging SNP	South African Black population			
		MAF: cases	MAF: controls	OR (95% CI)	P-value
rs2074356	-	0.0000	0.0000	NA	NA
rs13016963	-	0.3309	0.3333	0.99 (0.77-1.28)	0.9338
rs10931936, rs13016963, rs9288318, rs10201587	rs9288316	0.4694	0.4725	0.99 (0.78-1.26)	0.9187
rs4785204	rs12445755	0.1960	0.1961	1.00 (0.74-1.35)	0.9988
rs4785204	rs7204293	0.4802	0.4980	0.93 (0.73-1.18)	0.5609
rs7206735	rs6500291	0.2968	0.3137	0.92 (0.71-1.20)	0.5479
rs2239815	rs2097461	0.3849	0.3406	1.21 (0.94-1.56)	0.1332
rs2239815	rs5762788	0.3849	0.3406	1.21 (0.94-1.56)	0.1332
rs2239815	rs5762795	0.2248	0.1772	1.35 (0.99-1.82)	0.0531
rs2847281	-	0.0809	0.0824	0.98 (0.63-1.52)	0.9327
rs1042026	rs1229984	0.0324	0.0333	0.97 (0.49-1.90)	0.9300
rs1614972	-	0.3849	0.4216	0.86 (0.67-1.10)	0.2225
rs9257809	-	0.0054	0.0000	NA	0.0967
rs9257809	rs429479	0.0000	0.0000	NA	NA
rs9257809	rs406511	0.0054	0.0000	NA	NA
rs9257809	rs1233480	0.2428	0.2314	1.07 (0.80 - 1.41)	0.6612
rs9257809	rs442694	0.0000	0.0000	NA	NA
rs9257809	rs1535039	0.0000	0.0000	NA	NA
rs9257809	rs2746149	0.0755	0.0824	0.91 (0.58 - 1.42)	0.6800
rs9257809	rs2746150	0.0000	0.0000	NA	NA
rs9257809	rs1233491	0.0000	0.0000	NA	NA
rs9257809	rs2523443	0.0000	0.0000	NA	NA
rs9257809	rs404240	0.0000	0.0000	NA	NA
rs9257809	rs3749971	0.0018	0.0020	0.92 (0.06 - 14.70)	0.9512
rs9257809	rs3117427	0.0216	0.0235	0.92 (0.41 - 2.06)	0.8305
rs9257809	rs3117439	0.0324	0.0412	0.78 (0.41 - 1.48)	0.4445
rs9257809	rs3117425	0.0324	0.0412	0.78 (0.41 - 1.48)	0.4445
rs9936833	-	0.2356	0.2216	1.08 (0.81-1.44)	0.5856
rs9936833	rs1532167	0.2626	0.2627	1.00 (0.76-1.31)	0.9954
rs9936833	rs7186259	0.2500	0.2686	0.91 (0.69-1.19)	0.4880

5.5 Discussion

5.5.1 Case-control association analysis

After completing the case-control association study using the Immunochip, three SNPs in *TGFBR3* were significantly associated with OSCC using the Bonferroni correction to account for multiple testing ($P < 1.84 \times 10^{-6}$). Since this correction is conservative, additional variants with $P_{Immunochip} < 1 \times 10^{-4}$ were also selected for follow up. A well-powered independent replication study was not possible due to insufficient sample sizes. However, an extension study using an additional 126 cases and 577 controls provided improved power to investigate whether these variants were significantly associated with OSCC.

Seven variants were genotyped in this extension study, most of which were in regions with a high density of SNPs on the Immunochip. Several SNPs adjacent to the index variants also showed some evidence of association. These variants were prioritized over regions where only a single SNP showed association, as the latter were more likely to be genotyping artifacts. It should be noted that such regions may yet represent true associations. However, due to limited funds, not all SNPs that showed evidence of association could be genotyped in the extension study.

A case-control analysis of the extension study showed that the associations became less significant for six out of the seven SNPs, with one variant becoming more significant. However, none of the variants now showed a significant association using the Bonferroni correction as, although the sample sizes were larger, the observed effect sizes were reduced. More samples are needed to increase the power to establish whether these alleles affect OSCC susceptibility in the South African Black population. However, to achieve 80% power at a significance threshold of $P < 1.84 \times 10^{-6}$ and to detect the odds ratios that were observed in the extension study, the number of cases that would be needed ranges from ~500 to ~1,400 for the different variants with a larger number of controls also required. Sample collection is

continuing and additional recruitment centres are being established in other regions of South Africa.

Promisingly, some of the variants which showed evidence of association in the Immunochip study, and, thus included in the extension study, are located in or near genes which have a role in tumourigenesis. These will now be discussed.

5.5.1.1 *TGFBR3*

Four of the five SNPs most strongly associated with OSCC in the Immunochip analysis were all located in *TGFBR3*, *transforming growth factor beta receptor 3*. Three of these SNPs were the only variants to remain significant after correction for multiple testing. *TGFBR3* is located on chromosome 1p32 and encodes TβRIII (also known as betaglycan). TβRIII is a co-receptor, binding TGF-β superfamily ligands including TGF-β1-3, and bone morphogenetic proteins (BMPs) 2, 4 and 7, and facilitates their binding to TGF-β superfamily receptors (initially to type II receptors and then type I receptors) (reviewed in Gordon and Blobe 2008). Downstream, this leads to phosphorylation of the Smad family of transcription factors that regulate gene expression, as shown in Figure 5.15.

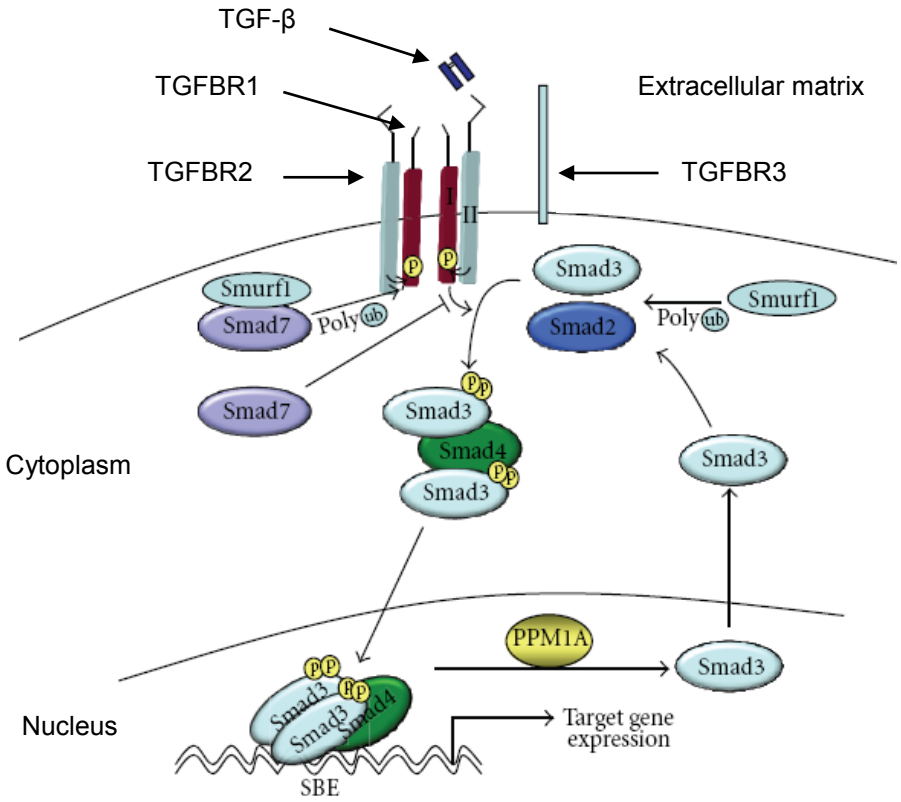


Figure 5.15: TGF-β signalling pathway
The ligand, TGF-β, binds to the TGF-β receptors (TGFBR1 and TGFBR2), facilitated by the co-receptor TGFBR3. This leads to phosphorylation and activation of Smad transcription factors which regulate gene expression. (Adapted from Glick 2012)

Defects in the TGF-β superfamily signaling pathways have been implicated in numerous diseases spanning a range of phenotypes including skeletal and muscular disorders, developmental disorders, cardiovascular diseases, as well as cancer development (reviewed in Gordon and Blobe 2008).

The TGF-β cytokine is involved in regulating a number of processes including differentiation, proliferation, angiogenesis, apoptosis as well as immune responses and embryonic development, and has a complex role in carcinogenesis (reviewed in Gatza *et al.* 2010). Early on in cancer formation, TGF-β acts as a tumour suppressor, inhibiting growth and inducing apoptosis. However, in later stages, the protein supports tumour progression by promoting processes such as angiogenesis, cell growth and survival, and invasion (Gordon and Blobe 2008; Bernabeu *et al.* 2009). This misregulation is caused by either the inactivation of critical components in the TGF-β signaling pathway such as the receptors and co-receptors, or by the loss of the tumour-suppressive part of the pathway (Massague 2008). This latter

process is due to the misregulation of other genes downstream in the cytostatic pathway which prevents the tumour-suppressor functions of TGF-β from acting but take advantage of its other regulatory functions to allow cancer cells to invade other regions and to evade immune responses (Massague 2008).

The chromosome arm containing *TGFBR3* frequently shows loss of heterozygosity (LOH) in several tumour types, including oesophageal, stomach, colon and rectum, and breast cancers (Ragnarsson *et al.* 1999). *TGFBR3* expression is also frequently downregulated in a number of cancers and correlates with a worsening disease progression including in breast, prostate, pancreatic adenocarcinoma and non-small cell lung cancer (Dong *et al.* 2007; Turley *et al.* 2007; Finger *et al.* 2008; Gordon *et al.* 2008). Increasing gene expression back to normal levels in both breast and prostate cancer cells inhibits some tumour-promoting activities (Dong *et al.* 2007; Turley *et al.* 2007). All of this evidence suggests that *TGFBR3* is a tumour suppressor gene and may be a good drug target in human cancers (Gatza *et al.* 2010).

Although expression of *TGFBR3* has not been investigated in OSCC, the *TGFBR2* receptor has been shown to play a role in the disease (Achyut *et al.* 2013). Achyut *et al.* first studied knock-out mice that harboured a *TGFBR2* deletion in stromal fibroblasts and found that inflammation and DNA damage was induced in adjacent epithelial cells, which led to the development of SCC of the mouse forestomach. The lining of the mouse forestomach is reported to be similar to that in the human oesophagus, which led to the authors to investigate *TGFBR2* in human OSCC. In stromal cells of OSCCs, reduced expression of *TGBFBRII* was observed, together with increased expression of inflammatory and DNA damaging mediators, which was suggested to drive human OSCC tumourigenesis (Achyut *et al.* 2013). This, therefore, shows that a functional TGF-β pathway is needed to prevent OSCC, and suggests that deregulation of other components of the pathway, including *TGRBR3*, may also result in the disease.

Polymorphisms in genes in the TGF-β pathway are associated with susceptibility to several cancers. For example, variants in *TGFBR1* are associated with overall cancer risk in a meta-analysis, which also includes significant associations with breast and ovarian cancer (Kaklamani *et al.* 2003). There have been no candidate gene association studies testing polymorphisms in *TGFBR3* with cancer susceptibility, and no associations have been identified in genome-wide association studies. Variants in *TGFBR3* have, however, been associated with a variety of other traits including bone mineral density (Xiong *et al.* 2009) and optic disc area (Khor *et al.* 2011). The function of *TGFBR3* makes it a plausible susceptibility gene in cancer development, and hence it is promising that four of the five top variants associated with OSCC in the Immunochip study are located in this gene.

Three of the four variants in *TGFBR3* that show evidence of association with OSCC are located in introns, with the fourth being downstream of the gene. None of these variants have previously been associated with any diseases or traits. However, several variants are predicted to be in regulatory regions. This was determined using RegulomeDB (http://regulome.stanford.edu/) (Boyle *et al.* 2012) and HaploReg (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) (Ward and Kellis 2012) which combine functional studies, such as ENCODE data (ENCODE Project Consortium 2011), and computational predictions into easy-to-use browsers. The SNP rs2810893 is predicted to be in a motif for the binding of the REST transcription factor. REST is suggested to be a tumour suppressor gene in colorectal cancer (Westbrook *et al.* 2005). The variant rs10493860 is predicted to be in a binding site for TRIM28 (KAP1) and SETDB1 which regulate transcription. TRIM28 is suggested to be a tumour suppressor in lung cancer (Chen *et al.* 2012.b), and the histone methyltransferase, *SETDB1*, is upregulated in a number of cancers, including lung, breast, ovarian and melanoma (Ceol *et al.* 2011). The variants rs9887787 and rs11165441 are not predicted to be in any regulatory regions. However, these variants may be in high LD with other SNPs which

are located in regulatory regions. Functional work may establish if any of these variants are involved in cancer development.

5.5.1.2 *PCSK9*

In the Immunochip analysis, rs2182833 is on the borderline of a significant association ($P_{Immunochip}$ = 1.85 x 10^{-6}) and is located upstream of *PCSK9*, *proprotein convertase subtilisin/kexin type 9*. The protein is involved in cholesterol metabolism and is able to regulate the number of low-density lipoprotein (LDL) receptors on the cell surface, hence controlling intracellular LDL levels (reviewed in Lopez 2008). Gain-of-function and loss-of-function mutations in *PCSK9* are associated with hypercholesterolemia and hypocholesterolemia, respectively. Some studies have suggested a modest association between low concentrations of LDL cholesterol in the plasma and an increased risk of cancer, although results are not convincing and have been conflicting (Law and Thompson 1991; Folsom *et al.* 2007; Strasak *et al.* 2009; Benn *et al.* 2011). The method by which LDL concentrations and cancer may be linked is unclear. It may be that low cholesterol levels contribute to cancer susceptibility, or cholesterol is lowered as an effect of cancer, or a confounder may effect both processes (Kritchevsky *et al.* 1991; Law and Thompson 1991). Nevertheless, this has led to the hypothesis that low levels of plasma LDL due to specific genotypes in *PCSK9* may also be associated with cancer susceptibility. However, two studies have not found any associations (Folsom *et al.* 2007; Benn *et al.* 2011).

This evidence does not support a role for *PCSK9* in cancer susceptibility and, hence, the association observed in the Immunochip study between rs2182833 and OSCC cannot be easily explained. In addition, the SNP is not predicted to be in any regulatory regions using HaploReg and RegulomeDB. Other genes in the vicinity of the variant include *BSND* (Bartter syndrome, infantile, with sensorineural deafness (Barttin), *TMEM61* (transmembrane protein 61) and *USP24* (ubiquitin specific peptidase 24). Only the latter has been linked to a potential role in carcinogenesis. *USP24* is a deubiquitinating enzyme, with polymorphisms in this gene associated with Parkinson's

disease (Li *et al.* 2006; Wu *et al.* 2010). Although no associations have been identified with *USP24* variants and cancer susceptibility, several other deubiquitinating enzymes are thought to play an important role in cancer development (Song *et al.* 2008; Pereg *et al.* 2010; Schwickart *et al.* 2010). This includes *USP9X* which is able to regulate protein levels of MCL1 by removing ubiquitin chains that would normally target the protein for degradation (Schwickart *et al.* 2010). *MCL1* is a member of the Bcl-2 protein family and promotes cell survival (Adams and Cory 1998), and overexpression of *MCL1* is observed in a variety of lymphomas and solid cancers (Kitada *et al.* 1998; Warr and Shore 2008). Other deubiquitinating enzymes, including *USP24,* may be found to have a role in cancer development.

5.5.1.3 *MTMR3*

Another gene of interest is *MTMR3*, *myotubularin related protein 3*, where seven SNPs in complete LD showed a suggestive association with OSCC in the Immunochip analysis ($P_{Immunochip}$ = 9.51 x 10^{-6}). All of the variants are located in introns, over a region of 60kb. *MTMR3* is a phosphoinositide 3-phosphate (PtdIns3P) phosphatase and is involved in the regulation of autophagy and in cell migration (Taguchi-Atarashi *et al.* 2010; Oppelt *et al.* 2012).

Autophagy is a mechanism to degrade and recycle or eradicate cellular components or pathogens in a response to stress or starvation (reviewed in Mathew *et al.* 2007). It is mainly controlled by the PI3 kinase pathway and its downstream protein ‘mammalian target of rapamycin’ (mTOR). The process has been shown to have a role in cancer development (reviewed in Mathew et al. 2007 and White *et al.* 2010). Under metabolic stress conditions, cells with defects in apoptosis (such as in tumour cells) had a sustained period of autophagy, which allowed the cells nutrient supply to be restored, enabling continued proliferation and growth. This is maintained until cells are unable to restore enough nutrients, when cell death eventually occurs.

The involvement of MTMR3 in autophagy is supported by evidence that the overexpression of *MTMR3* leads to the formation of small abnormal autophagsomes and a reduction in autophagic activity (Taguchi-Atarashi *et al.* 2010). In addition, *MTMR3* expression was downregulated in gastric cancer, which resulted in increased levels of autophagy (Lin *et al.* 2012). *MTMR3* expression was shown to be inhibited by a micro RNA, *miR-181a*, which itself was upregulated in gastric cancer tissue. *MTMR3* is predicted to contain two binding sites for miR-181a in the 3' UTR, one of which is only present when *MTMR3* rs12537 'C' allele is present. Hence, Lin *et al.* also performed a case-control association study for this variant in gastric cancer, and showed that individuals with CT genotypes at rs12537 had an increased risk of disease compared to those with CC genotypes ($P = 7.78 \times 10^{-6}$; OR = 1.86 (95% CI =1.42-2.45)). The variant rs12537 is included on the Immunochip but was not found to be associated with OSCC in the Black South African population, with minor allele frequencies of 46% and 49.4% in cases and controls, respectively (P = 0.327).

MTMR3, together with PIKfyve, a phosphoinositide 5-kinase, act on PtdIns3P to produce PtdIns5P (phosphatidylinositol 5-phosphate) which stimulates cell migration (Oppelt *et al.* 2012). MTMR3 has a key role in this, with depletion of MTMR3 protein decreasing cell migration (Oppelt *et al.* 2012). Cell migration is thought to be essential in cancer development, enabling cells to invade surrounding tissue (Yamaguchi *et al.* 2005), and hence, based on the findings of the Oppelt *et al.* study, higher MTMR3 levels would perhaps be expected in cancer development to lead to increased cell migration. This is not consistent with the studies described above where *MTMR3* levels were decreased in gastric cancer cells (Lin *et al.* 2012). However, MTMR3 depleted cells are less successful at migrating towards a wound, due to being unable to orientate their Golgi and cytoskeleton (Oppelt *et al.* 2012), which may prevent wound healing. This was discussed earlier as being a cause of inflammation which may lead to cancer development.

In addition to gastric cancer, variants in *MTMR3* have also been associated with childhood-onset inflammatory bowel disease (Henderson *et al.* 2011)

and lung cancer (Hu *et al.* 2011). In a lung cancer GWAS in a Han Chinese population, rs36600 was associated with the disease (Hu *et al.* 2011). The minor 'A' allele had frequencies of 11.9% in cases and 9.4% in controls, with an OR of 1.29 (95% CI = 1.20-1.38) and P = 6.2 x 10^{-13} in the combined GWAS and replication phases. This SNP showed a suggestive association ($P_{Immunochip}$ = 9.51 x 10^{-6}) with OSCC in our South African study, and a SNP in complete LD with it, rs36590, was genotyped in the extension study. The variant showed opposite effects in our study compared to the Chinese study, with the minor 'A' allele having a protective effect with minor allele frequencies of 2.6% and 6.2% in cases and controls, respectively (OR=0.41, 95% CI = 0.25 - 0.66; P = 1.6 x 10^{-4}). It is not unusual for variants to show opposing effects in different diseases or populations (this will be further explored in the Discussion, section 7.2). Hu *et al.* (2011) also note that their *MTMR3* region of association also extends to the neighboring *HORMAD2*, an open reading frame protein which may be involved in mitotic checkpoints and DNA repair. However, in our study, the region of association does appear to peak over the *MTMR3* gene (see Figure 5.10, p.166).

Apart from rs36600 which is associated with lung cancer, as described above, none of the SNPs in *MTMR3* showing evidence of association with OSCC in the South African population have been implicated in susceptibility to other diseases. Using RegulomeDB and HaploReg, some of the variants are predicted to be in regulatory regions. The variant rs36590 is predicted to alter the binding motif for POU2F1 (OCT-1) and AREB6 transcription factors; rs36600 is in a DNase hypersensitive site in several cell types; rs36596 and rs5752993 are in regulatory motifs for the CDX2 transcription factor binding site; rs5763634 is in a binding site for GATA1 and CEBPB transcription factors. The SNPs rs5763674 and rs4239932 are not predicted to be in any regulatory regions. All of these SNPs may also be in high LD with other variants that are predicted to be in regulatory regions.

5.5.1.4 Intergenic regions

The three other SNPs (rs13147507, rs13390918 and rs771403) that were genotyped in the extension study were all in intergenic regions. The nearest genes to rs13147507 are *ARSJ*, *UGT8* and *NDST4*. Of these genes, only *UGT8* (*UDP glycosyltransferase 8*) has been implicated in cancer, with high expression levels observed in metastatic prostate cancer cells compared to non-metaststic cells (Oudes *et al.* 2005). In addition, increased gene expression levels are associated with an increased risk of lung metastases in breast cancer patients (Landemaine *et al.* 2008). The variant is predicted to be located in binding sites for the transcription factors FOXA2 and CDP using RegulomeDB and HaploReg.

For rs13390918, the closest genes are *PLCL1* and *SATB2*. *PLCL1* is phospholipase C-like 1, and variants in *PLCL1* have previously been associated with Crohn's disease (Franke *et al.* 2010) and hip bone size (Liu *et al.* 2008) but its function remains unknown. Interestingly, variants in *PLCE1*, *phospholipase C epsilon 1*, have been associated with OSCC in Chinese populations (Abnet *et al.* 2010; Wang *et al.* 2010.a; Wu *et al.* 2011.c) and in our South African Black population (see Chapter 4). *SATB2*, *special AT-rich-binding protein 2*, is a transcription factor involved in several processes including the regulation of osteoblast differentiation and skeletal development (Dobreva *et al.* 2006). In addition, it may have a role in cancer development, with high expression levels observed in colorectal tumours (Magnusson *et al.* 2011; Eberhard *et al.* 2012). The variant is predicted to be in the binding motif for NANOG transcription factor using RegulomeDB.

The other intergenic variant is rs7714035, which is predicted (by RegulomeDB and HaploReg) to be in a regulatory region that binds several regulatory proteins including FOSL2, FOXA1, FOXA2, HNF4A, MAFF and MAFK. In addition, it is predicted to alter the POU1F1 transcription factor binding site. The variant is closest to *C5orf30*, *GIN1* (*gypsy retrotransposon integrase 1*), *NUDT12* (*nudix (nucleoside diphosphate linked moiety X)-type motif 12*) and PAM (*peptidylglycine alpha-amidating monooxygenase*), none

of which have a known function that is likely to contribute to cancer development. *PPIP5K2* (*diphosphoinositol pentakisphosphate kinase 2*) is also in the region, and is involved in cell signaling, although the exact function of the gene is yet to be determined. It is known that *PPIP5K2* (and *PPIP5K1*) binds PtdIns(3,4,5)P3 with a high affinity which may prevent other ligands from binding or displace weaker-bound ligands (Gokhale *et al.* 2011). PtdIns(3,4,5)P3 is involved in the phosphatidylinositol 3-kinase (PI3K) pathway, and phosphorylation of PtdIns(3,4,5)P3 by PI3K leads to the activation of the kinase AKT which regulates processes including cell growth, proliferation and apoptosis (Vivanco and Sawyers 2002). Mutations in several genes in the PI3K pathway have been identified in a variety of cancers (reviewed in Vivanco and Sawyers 2002). Therefore, the potential ability of *PPIP5K2* to regulate PtdIns(3,4,5)P3 levels through its binding may be important in carcinogenesis.

Taken together, there is limited evidence that these three intergenic variants are involved with cancer development. All SNPs are near genes that may be involved in tumourigenesis, including *UGT8, SATB2* and *PPIP5K2*, but there is no evidence that the variants affect the expression of these genes. However, all three variants are in regulatory regions, which will require functional studies to establish whether they have a role in cancer development.

5.5.1.5 Gene-environment interactions

Only one variant, rs2182833, in the extension study showed evidence of gene-environment interactions. Unusually, the frequency of this SNP in controls differed between smokers and non-smokers (0.289 and 0.368, respectively) but had similar frequencies in smoker and non-smoker cases (0.418 and 0.401, respectively). This resulted in a significant association of the variant in smokers ($P = 7.0 \times 10^{-6}$) but not in non-smokers ($P = 0.2825$). The difference in minor allele frequencies between non-smoking cases and controls is difficult to interpret. Our expectation for gene-environmental interactions is that the minor allele frequency will differ in cases, showing

that, for example, a higher frequency of the risk allele in smokers compared to non-smokers increases the risk of disease in smokers. The number of smokers and non-smokers in controls in this study is 327 and 496, respectively, and the SNP is common, indicating that the results observed are not due to a lack of rare homozygotes or heterozygotes. Additionally, this SNP has not been associated with smoking behavior in GWAS so there is no prior expectation for frequencies to differ in controls. Expansion of sample sizes for both cases and controls would help to establish whether this is a robust finding. The lack of interactions with alcohol drinking is not surprising considering none of the variants have a known role in alcohol metabolism.

In addition to this work, an Immunochip-wide gene-environmental interaction test was performed, which used logistic regression adjusted for age, gender, alcohol consumption and tobacco smoking. This approach was based on a study by Wu *et al.* (2012.a) who performed a genome-wide gene-environmental interaction test and were the first to analyze OSCC GWAS data in this manner. This led to the identification of 15 variants located at the *ALDH2* locus on chromosome 12q24, which were associated with OSCC in the Chinese population. SNPs at this locus had previously been found to interact with alcohol to affect disease susceptibility (Wu *et al.* 2011.c). Two novel susceptibility loci, rs9288520 and rs17450420 on chromosome 2q22 and 13q33, respectively, were also associated with OSCC. A stratified analysis, whereby case-control studies for drinkers and non-drinkers were performed separately, showed that the minor alleles of both variants were associated with a decreased disease risk in non-drinkers but an increased risk in drinkers (Wu *et al.* 2012.a). These SNPs were not associated in the GWAS alone, without an interaction test. This, therefore, suggests the importance of considering environmental risk factors at an early stage of analysis rather than stratifying only significant case-control associations by these factors, as is the standard approach to analysis.

Analysis of the South African Immunochip-wide gene-environmental interaction test did not identify any variants associated with OSCC that would survive the Bonferroni correction ($P<1.84 \times 10^{-6}$). In the gene-alcohol

interaction test, the most strongly associated SNPs were rs10493941 and rs12726628 (both with P=1.83 x 10^{-4}) in *SLC30A7* (*solute family carrier 30 (zinc transporter), member 7*). An additional four of the top 10 variants were also in this region, which also included *DPH5* downstream of *SLC30A7*. *SLC30A7* is a zinc transporter and does not have any known roles in alcohol metabolism, although another solute family carrier (*SLC10A2*) was associated with OSCC in the gene-alcohol interaction test by Wu *et al*. (2012.a). In the gene-smoking interaction test, rs1052240 in the intron of *PTPRC* was the most strongly associated SNP (P = 1.66 x 10^{-5}). *PTPRC* (*protein tyrosine phosphatase, receptor type, C*), also known as CD45, is involved in the regulation of immune cell signaling (Hermiston *et al.* 2003; Saunders and Johnson 2010). It is also involved in cancer development, acting as a tumour suppressor in T-cell acute lymphoblastic leukemia (Porcu *et al.* 2012). As none of these variants had a significant association with OSCC, a stratified analysis was not performed.

None of the 20 variants identified through the Immunochip-wide gene-environmental interaction tests were associated with OSCC in the case-control study (P>0.2). Since these variants do not show evidence of involvement in alcohol or smoking pathways, and fail to meet the Bonferroni correction for a significant association, it is unlikely they play a role in OSCC susceptibility in the South African Black population.

5.5.2 Population structure

Principal components analysis was used to investigate population structure in the South African samples. The Mixed Ancestry population showed a large degree of heterogeneity, which is consistent with historical knowledge of the population being formed from indigenous Khoi, Europeans, Asians and other African populations. A previous large-scale analysis of the Mixed Ancestry population by de Wit *et al*. (2010) estimated that the main ancestral groups were Khoesan (32%-43%), Bantu-speaking Africans (20-36%), Europeans (21-28%) and Asians (9-11%).

In contrast to the Mixed Ancestry population, the South African Black cases and controls are tightly clustered in the PCA plots, forming a distinct group indicative of a relatively homogeneous genetic architecture. This is consistent with the fact that almost all patients and controls are Xhosa-speakers, originating from a single linguistic ancestral population. Interestingly, the Black individuals clustered adjacent to the HapMap Yoruban subjects, consistent with a common sub-Saharan Bantu-speaking ancestry (Tishkoff *et al.* 2009). The only other study of an extensive panel of genetic markers in the Xhosa population (also known as isiXhosa) is from Patterson *et al.* (2010) who genotyped 20 individuals on an Affymetrix 900K SNP array. This also showed tight clustering of the Xhosa in the PCA plots, closely adjacent to the Yoruban population.

5.5.3 Replication of Chinese GWAS hits using the South African Immunochip data

The use of the Immunochip containing ~200,000 variants also allowed other SNPs of interest to be investigated, such as those associated with OSCC in other populations. For the 40 variants identified through Chinese GWAS and the European Barrett's oesophagus GWAS, only 6 were present on the Immunochip, with an additional 7 covered by proxies. This low number is probably due to the majority of the SNPs not having a known role in immune-related diseases, for which the Immunochip array was designed. None of the index SNPs or proxies were associated with OSCC in the Black population, all with $P>0.05$. These loci may represent population-specific disease associations. Alternatively, different variants in the region may be associated in the South African population but the effect could not detected by genotyping of the index SNP. This will be discussed in the following section. Additionally, our study had low power to detect modest genetic effect sizes, with only 278 cases and 257 controls tested for disease association. This compares to the thousands of samples included in the GWAS studies (Abnet *et al.* 2010; Wang *et al.* 2010.a; Wu *et al.* 2012.a). A larger sample set would

help to determine whether the Chinese GWAS hits also contribute to OSCC susceptibility in the South African Black population.

5.5.4 Summary of case-control study

This study found three variants that were significantly associated with OSCC in the South African Black population using the Immunochip. These SNPs were all located in *TGFBR3*, which is a strong candidate gene for cancer development due to its involvement in the regulation of processes such as proliferation, angiogenesis, apoptosis and immune responses. Other variants with a suggestive association were also in regions that have previously been associated with carcinogenesis, including *MTMR3*.

Following an extension study which genotyped 7 variants in an expanded set of cases and controls, no variants were significantly associated with OSCC. The significance threshold ($P<1.84 \times 10^{-6}$) was determined using the Bonferroni correction for the number of independent SNPs tested on the Immunochip. However, this is a conservative correction, which the top SNP in *TGFBR3* just failed to meet ($P= 4.0 \times 10^{-6}$). In addition, sample numbers were relatively small with 407 cases and 834 controls.

It is possible that these loci are associated with OSCC in the South African population but a significant association could not be detected. This may be due to several factors. Firstly, and perhaps most importantly, the chip was designed for European populations. Therefore, if an index SNP was merely a tagging SNP in a high level of LD with the causal variant in a European population, then an association may not be observed in African populations. This is due to Africans having a lower level of LD than Asian and European populations (Teo *et al.* 2010). If the index SNP was in moderate LD with the causal variant in an African population, then a significant disease association may still not be observed due to a loss of power to detect a weaker effect size (Teo *et al.* 2010). Additionally, an array designed for Europeans results in the absence of SNPs (both common and rare) which are only present in African populations. Thus, the variants in regions which show evidence of

association with OSCC in the South African Black population may not have been fully explored. The Immunochip was also designed several years ago and more variants in these loci are likely to have been identified which were not included. Further follow up of the results from the Immunochip scan should be carried out when a larger number of cases and controls from this population are available to increase the power of the study.

Alternatively, the results may indicate that none of the regions are significantly associated with OSCC in the South African Black population. Several SNPs on the Immunochip have previously been associated with OSCC in Chinese populations (either directly or as tagging SNPs) (Wu *et al.* 2011.c; Abnet *et al.* 2012; Wu *et al.* 2012.a), suggesting that these regions may be important in disease susceptibility. Therefore, these loci may represent population-specific disease associations, with the risk loci absent in the South African population, but a much larger sample size would be required to exclude them. Alternatively, gene-environment interactions may be important, with environmental risk factors being population-specific.

6 Identification of somatic mutations in oesophageal squamous cell carcinoma

6.1 Known somatic mutations in OSCC

Somatic mutations that occur in OSCC have been explored for over 20 years. In 1990, the first mutations to be discovered were in *TP53,* which were found using a candidate gene sequencing approach (Hollstein *et al.* 1990). In that study, *TP53* mutations were identified in 2/4 (50%) of OSCC cell lines and 5/14 (36%) of OSCC tumours. In the years since, *TP53* has been investigated in OSCC patients in different in populations, with mutation rates ranging from 17-84% (reviewed in Egashira *et al.* 2007). The candidate gene approach has also been successful in identifying somatic mutations in several other genes including *MTS1*/*CDH4I* (52% of tumours) (Mori *et al.* 1994), *PIK3CA* (12%) (Phillips *et al.* 2006; Maeng *et al.* 2012), *p16*/*CDKN2* (28%) (Gamieldien *et al.* 1998) and *MLH1* (8%) (Maeng *et al.* 2012).

In more recent years, technologies have been developed enabling genome-wide mutation analysis to investigate all somatic mutations present in a tumour. The first and only study to perform this in OSCC was by Agrawal *et al.* (2012), who sequenced the exome of 12 OSCC tumours and matched normal tissue from patients in the USA. The tumours contained an average of 83 somatic mutations, with the most frequently mutated gene being *TP53*, present in 92% of tumours. Other frequently mutated genes (≥3/12 tumours) were *NOTCH1*, *NOTCH3*, *FBXW7, KIF16B, KIF21B* and *MYCBP2*. These genes, together with *NOTCH2*, were sequenced in an additional 41 tumours and normal tissues and resulted in the following mutation frequencies: *TP53* (62% of tumours), *NOTCH1* (21%), *NOTCH2* (6%), *NOTCH3 (8%)*, *FBXW7 (6%),* with no mutations found in *KIF16B, KIF21B* and *MYCBP2*. Agrawal *et al.* also sequenced these five frequently mutated genes in 48 OSCC samples and matched controls from a Chinese population. Of these, 71% harboured a mutation in *TP53* but *NOTCH* mutations were rare, occurring in 3 tumours (one mutation in each *NOTCH1*, *NOTCH2* and *NOTCH3*). The authors conclude that the mechanisms of tumourigenesis may vary between different populations which could have implications for the success of drug treatments

in populations where the mutational landscape is not known. Interestingly, *TP53*, *NOTCH1*, *NOTCH2*, *NOTCH3*, *FBXW7* are also mutated in head and neck squamous cell carcinoma (Agrawal *et al.* 2011; Stransky *et al.* 2011).

A summary of somatic mutations occurring in OSCC is shown in Table 6.1, the majority of which were identified using candidate gene studies. As many *TP53* studies have been performed, a review article has been included which summarizes these studies (Egashira *et al.* 2007).

Table 6.1: Summary of somatic mutations in OSCC

Study	Study type	Gene	Number of samples mutated	% samples mutated
Agrawal *et al.* 2012	Exome sequencing - discovery phase	*TP53*	11/12	92%
		NOTCH1	4/12	33%
		NOTCH3	3/12	25%
		FBXW7	2/12	17%
	Exome sequencing - follow-up	*TP53*	-	62%
		NOTCH1	-	21%
		NOTCH2	-	6%
		NOTCH3	-	8%
		FBXW7	-	6%
Maeng *et al.* 2012	Candidate gene	*PIK3CA*	10/80	11.5%
		MLH1	7/80	8.0%
		TP53	3/80	3.5%
		BRAF	1/80	1.2%
		CTNNB1	1/80	1.2%
		EGFR	1/80	1.2%
Egashira *et al.* 2007	Candidate gene	*TP53*	45/95	47.4%
	Review of *TP53* mutation studies		-	17% - 84%
Phillips *et al.* 2006	Candidate gene	*PIK3CA*	4/35	11.8%
		PIK3CB	0/35	0.0%
Hu *et al.* 2004	Candidate gene	*CDKN2A*	14/56	25%
		CDKN2B	1/56	1.8%
Li *et al.* 2003	Candidate gene	*DICE*	3/56	5.4%
Lo *et al.* 2002	Candidate gene	*RNF6*	3/24	12.5%
Giroux *et al.* 2002	Candidate gene	*CDKN2A*	6/100	6.0%
Gamieldien *et al.* 1998	Candidate gene	*CDKN2A*	21/76	28%
		TP53	13/76	17%
Esteve *et al.* 1996	Candidate gene	*CDKN2A*	2/21	9.5%
Suzaki *et al.* 1995	Candidate gene	*CDKN2A*	5/35	14%
		CDKN2B	1/39	3%
Mori *et al.* 1994	Candidate gene	*MTS1/CDH4I*	14/27	51.9%

6.2 Somatic mutations in OSCC from South African populations

Only one study has analyzed OSCC somatic mutations in South African patients (Gamieldien *et al.* 1998). This used a candidate gene approach and focused on *TP53* exons 5-8 and *p16*/*CDKN2* exons 1-2, and found that these genes were mutated in 17% and 28% of tumours, respectively.

The aim of this chapter is to further explore the somatic changes occurring in OSCC patients from the South African population using a whole-exome sequencing approach. This will be achieved by initially sequencing matched blood and tumour DNA from 8 OSCC patients in order to identify driver mutations.

6.3 Exome sequencing of OSCC blood-tumour pairs

6.3.1 Exome sequencing metrics

Eight blood-tumour pairs were whole-exome sequenced and the summary statistics are shown in Table 6.2. The median sequencing depth ranged from 52-283x. The samples sequenced by Illumina (232T and P662) had the highest median depths (283 and 244, respectively), with those sequenced at KCL having higher values than ICR (71-129 vs. 52-66, respectively). The percentage of targeted regions covered at 40x ranged from 63-93%.

Table 6.2: Summary statistics for whole-exome sequencing

Sample	Sequencing location	Total reads	% on-target	Percentage of targeted regions covered at:			Median depth
				10x	20x	40x	
T416	ICR	94061126	72	93	87	71	63
P1354	ICR	98554907	70	93	87	71	64
T438	ICR	96023847	74	94	87	72	66
P1400	ICR	82668200	67	93	85	64	52
T441	ICR	90426864	74	94	87	71	62
P1116	ICR	83906875	68	92	84	63	52
T442	ICR	97546354	72	93	87	72	65
P1406	ICR	78138522	70	93	86	64	52
T443	ICR	94027141	71	93	86	69	61
P1408	ICR	90488711	71	93	86	69	60
386T	KCL	110579550	78	95	92	85	125
P1282	KCL	114420987	77	96	92	86	129
T437	KCL	61513647	78	93	87	74	71
P1377	KCL	108294987	80	96	92	86	128
232T	KCL	270262199	70	98	96	93	283
P662	KCL	240263017	68	98	95	91	244

6.3.2 Somatic mutations identified

Several thresholds were applied for a somatic mutation to be called. Firstly, the total number of sequencing reads required was ≥8 and ≥14 for blood and tumour sample, respectively. In addition, the threshold for the frequency of the mutant alternative allele was set at ≥15% in the tumour and <2% in the normal blood DNA. However, the alternative allele frequency was lowered to ≥10% in two tumours in which very few mutations were identified using the ≥15% threshold.

The number of potential somatic mutations in each of the blood-tumour pairs is summarized in Table 6.3, with the results visualised in Figure 6.1.

Table 6.3: Summary of potential somatic mutations
Mutations are present in ≥15% of sequencing reads in the tumour and absent (<2%) in the blood DNA (**a**). The threshold was lowered to ≥10% of sequencing reads for samples with a low number of mutations (**b**).

a)

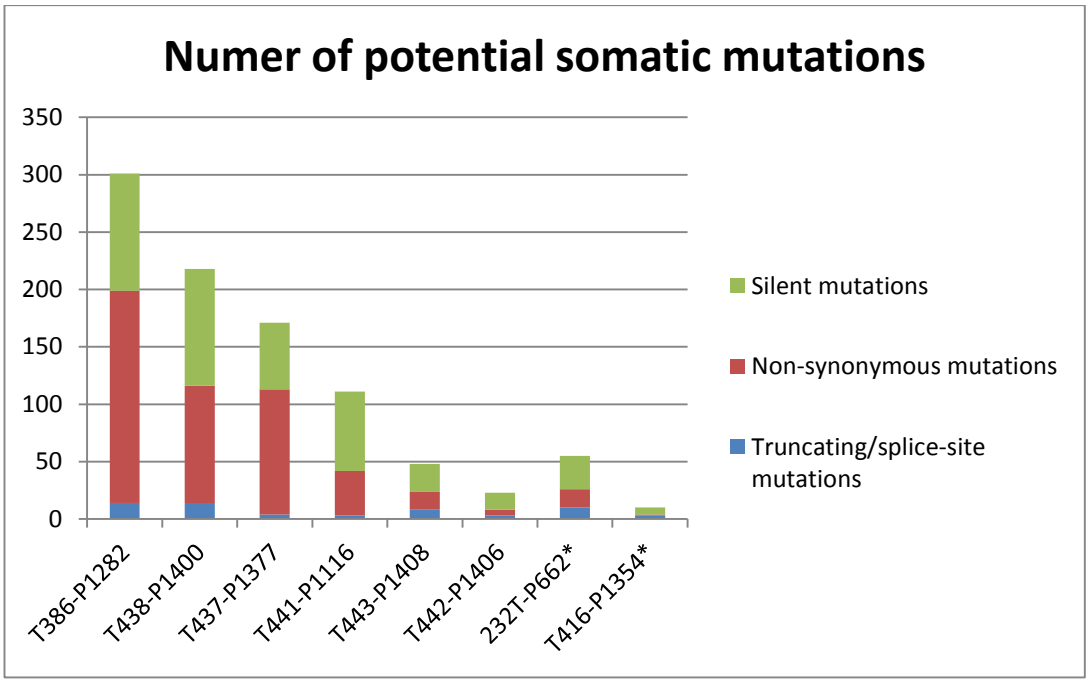
Sample	Total number of somatic mutations (x)	Truncating/ splice-site mutations (y)	Non-synonymous mutations	Silent mutations (z)	Number of truncating/splice-site mutations and non-synonymous mutations in cancer gene census genes
T386-P1282	301	14	185	102	8
T438-P1400	215	13	103	102	8
T437-P1377	171	4	109	58	7
T441-P1116	111	3	39	69	2
T443-P1408	48	8	16	24	0
T442-P1406	23	3	5	15	0
232T-P662	6	1	2	3	0
T416-P1354	3	0	1	2	0

b)

Tumour-blood pair	Total number of somatic mutations (x)	Truncating/ splice-site mutations (y)	Non-synonymous mutations	Silent mutations (z)	Number of truncating/splice-site mutations and non-synonymous mutations in cancer gene census genes
232T-P662	55	10	16	29	0
T416-P1354	10	1	3	6	0

x = Stop-gain/loss, frameshift, non-synonymous, splice sites, synonymous, intergenic, intronic, UTR
y = Stop-gain/loss, frameshift, essential splice sites
z = Synonymous, intergenic, intronic, UTR, upstream, non-essential splice sites

a)



b)

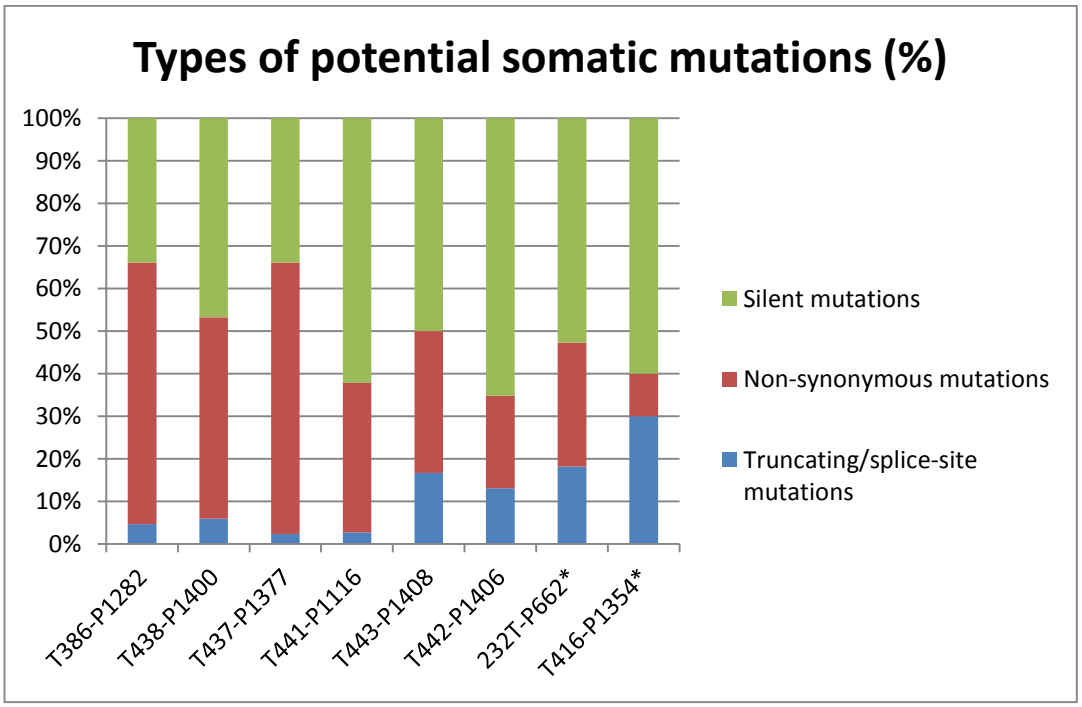


Figure 6.1: Summary of potential somatic mutations
The number **(a)** and percentage **(b)** of each type of somatic mutation present in the blood-tumour pairs. The minor allele frequency (MAF) for each variant was ≥15% in the tumours, apart from 232T and T416 which used ≥10% MAF (indicated by *).

The total number of potential somatic mutations was variable between tumours, ranging from 10 to 301. The number of protein truncating/splice-site mutations (stop-gain/loss, frameshift insertions or deletions, and essential splice sites) ranged from 1 to 14 in each tumour. However, not all mutations appeared to be real when the sequencing data was visualised using Integrated Genomics Viewer (IGV). IGV was used to view stop-gain/loss, frameshift and essential splice sites mutations, together with non-synonymous mutations that were in genes present in the Cancer Gene Census, in all blood-tumour pairs. These mutations were considered more likely to be involved in cancer development, and hence were prioritised over silent mutations. Table 6.4 shows the potentially functional somatic mutations identified, together with a prediction of whether they appear to be valid mutations on IGV.

Table 6.4: Somatic mutations identified by whole-exome sequencing
The reference (ref) and alternative (alt) alleles are shown, together with the minor allele frequency (MAF) for the alternative allele in the tumour. All tumours used a threshold of ≥15% of reads supporting the alternative allele, apart from 232T-P662 and T416-P1354 which used a ≥10% threshold. Each mutation was visualised using Integrated Genomics Viewer (IGV) to determine whether they appeared to be valid mutations.

Sample	Gene	Position (Hg19)	Ref allele	Alt. allele	MAF in tumour	Effect	Valid on IGV
	C9orf142	9:139887697	T	G	0.17	Essential splice site	No
	CRHBP	5:76249852	A	C	0.26	Essential splice site	No
	DPP6	7:154681166	G	A	0.16	Essential splice site	No
	ERCC5	13:103498666	T	G	0.22	Non-synonymous	No
	FANCA	16:89805074	C	G	0.17	Non-synonymous	No
	GNAS	20:57429620	G	C	0.17	Non-synonymous	No
	GNAS	20:57429907	A	C	0.21	Non-synonymous	No
T386-P1282	*IQCH*	15:67547287	T	G	0.16	Essential splice site	No
	KNSTRN	15:40675525	T	G	0.16	Essential splice site	No
	MEF2C	**5:88026048**	**C**	**A**	**0.16**	**Stopgain**	**Yes**
	MKL1	**22:40815208**	**C**	**A**	**0.16**	**Non-synonymous**	**Yes**
	MN1	22:28194178	T	G	0.17	Non-synonymous	No
	NOXO1	16:2030366	A	C	0.15	Essential splice site	No
	PPM1D	**17:58740836**	**C**	**T**	**0.23**	**Stopgain**	**Yes**
	PPP2R1A	19:52723071	T	G	0.18	Non-synonymous	No
	PRELID1	5:176733533	G	C	0.22	Stoploss	No

	RBM26	**13:79940771**	**-**	**T**	**0.21**	**Frameshift**	**Yes**
	TP53	**17:7579389**	**G**	**A**	**0.31**	**Stopgain**	**Yes**
	TSC22D2	3:150127930	-	C	0.15	Frameshift	No
	TTC22	1:55251654	A	C	0.15	Essential splice site	No
	ZZEF1	17:4045834	A	C	0.17	Essential splice site	No
T438-P1400	***APC***	**5:112173713**	**G**	**C**	**0.30**	**Non-synonymous**	**Yes**
	APC	**5:112174228**	**G**	**A**	**0.27**	**Non-synonymous**	**Yes**
	APC	**5:112174347**	**G**	**C**	**0.21**	**Non-synonymous**	**Yes**
	ARHGAP21	**10:24884912**	**T**	**TA**	**0.19**	**Frameshift**	**Yes**
	ATAD5	**17:29214214**	**AT**	**A**	**0.30**	**Frameshift**	**Yes**
	CARS	**11:3059285**	**C**	**G**	**0.24**	**Non-synonymous**	**Yes**
	CNGB1	**16:57996877**	**C**	**T**	**0.45**	Essential splice site	**Yes**
	CORO2B	**15:68937546**	**C**	**A**	**0.27**	**Stopgain**	**Yes**
	FCRL3	**1:157667452**	**G**	**A**	**0.23**	**Stopgain**	**Yes**
	KRT27	**17:38937523**	**C**	**G**	**0.35**	Essential splice site	**Yes**
	MLL2	**12:49448371**	**C**	**T**	**0.26**	**Non-synonymous**	**Yes**
	NLRC5	**16:57111288**	**TAG**	**T**	**0.32**	**Frameshift**	**Yes**
	NOTCH2	**1:120512275**	**C**	**CA**	**0.17**	**Frameshift**	**Yes**
	OR52A5	**11:5153238**	**AACCCT**	**A**	**0.17**	**Frameshift**	**Yes**
	OVGP1	1:111957563	C	CT	0.22	Frameshift	No
	PMS1	**2:190738302**	**G**	**C**	**0.24**	**Non-synonymous**	**Yes**
	SP1	**12:53777373**	**C**	**T**	**0.33**	**Stopgain**	**Yes**
	TP53	**17:7578466**	**G**	**A**	**0.36**	**Non-synonymous**	**Yes**
	TP53	**17:7578458**	**G**	**GGA**	**0.30**	**Frameshift**	**Yes**
	ZNF521	**18:22804986**	**C**	**G**	**0.19**	**Non-synonymous**	**Yes**
	ZNF750	**17:80789692**	**G**	**GA**	**0.28**	**Frameshift**	**Yes**
T437-P1377	*DGKI*	7:137282649	C	A	0.17	Stopgain	No
	FCGR2B	1:161642797	T	G	0.18	Non-synonymous	No
	HOXA11	7:27224414	T	G	0.15	Non-synonymous	No
	MLL2	12:49426251	T	G	0.19	Non-synonymous	No
	MLLT6	17:36861953	T	G	0.18	Non-synonymous	No
	RECQL4	**8:145737856**	**G**	**T**	**0.16**	**Non-synonymous**	**Yes**
	RECQL4	8:145738086	A	G	0.19	Non-synonymous	No
	SCARB1	12:125299662	T	G	0.16	Essential splice site	No
	SCN8A	12:52115555	-	A	0.17	Frameshift	No
	TAF15	17:34171676	A	G	0.18	Non-synonymous	No
	TARBP1	1:234601455	C	A	0.20	Essential splice site	No
T441-P1116	***ARHGEF2***	**1:155931616**	**TGATAAATACCC**	**T**	**0.16**	**Frameshift**	**Yes**

	COX6C	**8:100899805**	**G**	**C**	**0.20**	**Non-synonymous**	**Yes**
	FLT3	**13:28611364**	**C**	**G**	**0.17**	**Non-synonymous**	**Yes**
	FZD6	**8:104340628**	**C**	**T**	**0.15**	**Stopgain**	**Yes**
	IL21R	**16:27459982**	**C**	**T**	**0.23**	**Non-synonymous**	**Yes**
	LARGE	**22:33780177**	**C**	**T**	**0.32**	Essential splice site	**Yes**
	PAX7	**1:18961022**	**G**	**A**	**0.16**	**Non-synonymous**	**Yes**
	TET2	**4:106156540**	**C**	**T**	**0.32**	**Stopgain**	**Yes**
	TMPRSS2	**21:42843880**	**C**	**T**	**0.32**	**Non-synonymous**	**Yes**
T443-P1408	***DNAH10***	**12:124285943**	**AG**	**A**	**0.18**	**Frameshift**	**Yes**
	JUNB	**19:12902663**	**C**	**CCT**	**0.33**	**Frameshift**	**Yes**
	KL	**13:33628178**	**AC**	**A**	**0.34**	**Frameshift**	**Yes**
	OSBPL3	**7:24901235**	**T**	**A**	**0.32**	**Stopgain**	**Yes**
	RORA	**15:60789688**	**G**	**T**	**0.16**	**Stopgain**	**Yes**
	RTL1	**14:101348698**	**G**	**A**	**0.17**	**Stopgain**	**Yes**
	SRPX	**X:38080687**	**CAA**	**C**	**0.33**	**Frameshift**	**Yes**
	VAMP4	**1:171678834**	**G**	**A**	**0.21**	**Stopgain**	**Yes**
T442-P1406	***GPRASP2***	**X:101970164**	**A**	**T**	**0.15**	**Stopgain**	**Yes**
	HPX	**11:6458700**	**G**	**A**	**0.18**	**Stopgain**	**Yes**
	WDR17	**4:177083305**	**CT**	**C**	**0.18**	**Frameshift**	**Yes**
232T-P662	*BBC3*	19:47729819	A	C	0.10	Essential splice site	No
	DPP7	9:140007195	A	C	0.11	Essential splice site	No
	DPY19L1	7:34979763	A	C	0.13	Essential splice site	No
	FAM131C	**1:16384994**	**C**	**-**	**0.10**	**Frameshift**	**Yes**
	MTR	1:237038025	G	A	0.11	Essential splice site	No
	MYH14	**19:50789941**	**C**	**-**	**0.10**	**Frameshift**	**Yes**
	NFIA	1:61872232	A	C	0.13	Essential splice site	No
	NFIA	1:61872233	G	C	0.17	Essential splice site	No
	SREBF1	17:17716680	A	C	0.11	Essential splice site	No
	TOPAZ1 / C3orf77	**3:44283603**	**C**	**T**	**0.10**	**Stopgain**	**Yes**
T416-P1354	***CXorf30***	**X:36324916**	**AT**	**A**	**0.13**	**Frameshift**	**Yes**

Examples of mutations that appear to be valid and those that are unconvincing using IGV are shown in Figure 6.2 and Figure 6.3, respectively. The former shows the mutation to be absent in the blood and present in a number of reads in the tumour. The latter places the identified mutation in a

region containing many variants, both in the blood and tumour. The degree of confidence in calling these variants is varied, with the lighter coloured letters indicating a lower confidence. This may represent mis-alignment or sequencing errors.

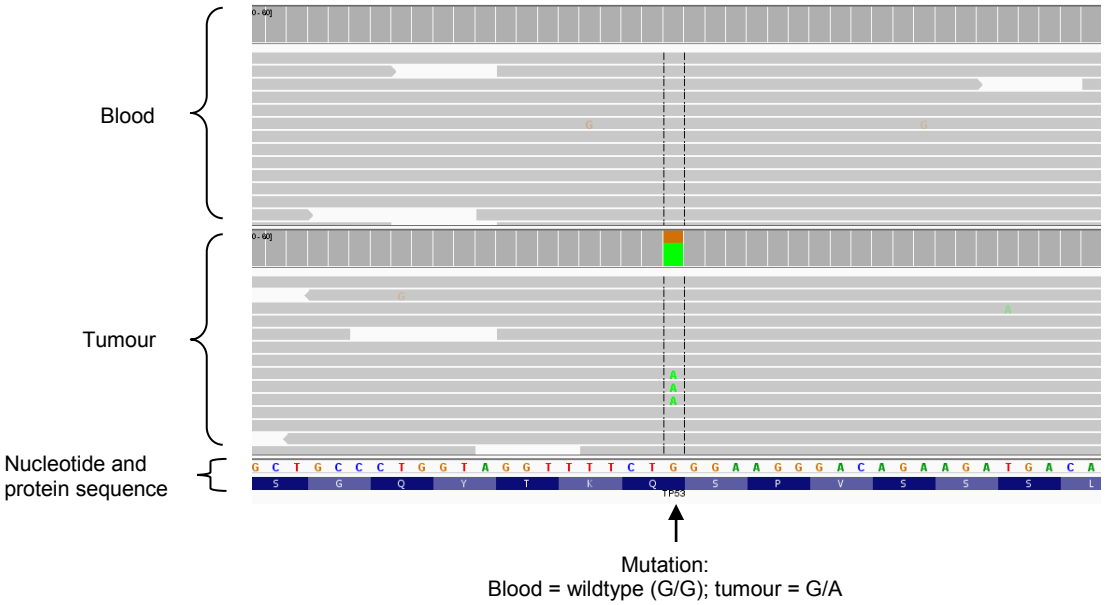


Figure 6.2: Sequencing reads for a valid somatic mutation
This mutation was identified by whole-exome sequencing and visualised using IGV. The mutation is clearly present in the tumour and absent in the blood.

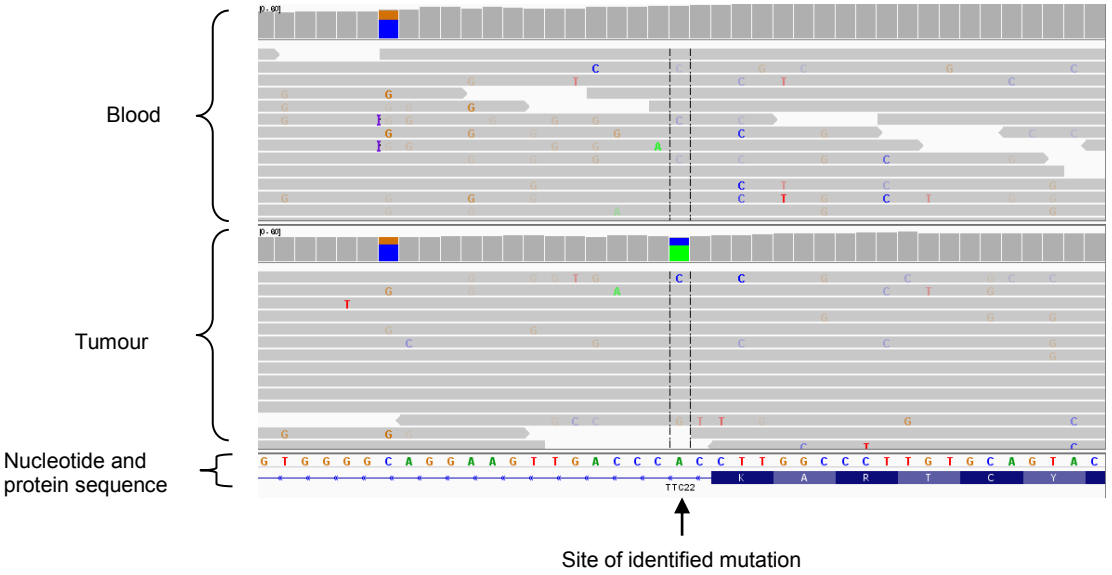


Figure 6.3: Sequencing reads for an unconvincing somatic mutation
This mutation was identified by whole-exome sequencing but by visualising the region on IGV, it does not appear to be valid.

The number of mutations that were confirmed on IGV is summarized in Table 6.5 for each blood-tumour pair.

Table 6.5: Number of somatic mutations confirmed on IGV

Sample	Mutations confirmed on IGV (%)
T386-P1282	5/21 (24%)
T438-P1400	20/21 (95%)
T437-P1377	1/11 (9%)
T441-P1116	9/9 (100%)
T443-P1408	8/8 (100%)
T442-P1406	3/3 (100%)
232T-P662	3/10 (30%)
T416-P1354	1/1 (100%)

Five blood-tumour pairs had a high confirmation rate (>95%). However, three pairs had low rates (<30%), which were the samples where the mutation calling was carried out at KCL, whereas the others were analysed at the ICR. To ensure that these results did not reflect a poor performing calling algorithm, all other blood-tumour pairs were run through the KCL analysis pipeline, with results shown in Table 6.6.

Table 6.6: Comparison of mutations confirmed on IGV using the ICR and KCL analysis pipelines

Sample	Mutations confirmed on IGV (%)	
	ICR	KCL
T438-P1400	20/21 (95%)	16/17 (94%)
T441-P1116	9/9 (100%)	9/14 (64%)
T443-P1408	8/8 (100%)	8/10 (80%)
T442-P1406	3/3 (100%)	6/9 (67%)
T416-P1354	1/1 (100%)	9/23 (39%)

The percent of mutations confirmed using the KCL calling algorithm is lower than that for the ICR method (39-94% versus 95-100%, respectively). This suggests that the KCL algorithm could be further optimized. However, in 4 out of 5 tumours, the KCL method identified more mutations than ICR, with a similar number of mutations confirmed (apart from T416-P1354), although they were not always the same mutations. Therefore, the KCL method may introduce more false positives but is still be able to identify true positives.

Hence, the specificity of the KCL method is lower, but the sensitivity is higher.

Mutations which appeared to be valid on IGV were selected for Sanger sequencing to confirm their presence using an independent technique.

6.3.3 Sanger sequencing to confirm somatic mutations

Protein truncating/splice-site mutations and non-synonymous mutations that were in genes present in the Cancer Gene Census were prioritized for confirmation by Sanger sequencing. The variants selected for Sanger sequencing are shown in Table 6.7, together with the results.

Table 6.7: Confirmation of somatic mutations using Sanger sequencing

Sample	Gene	Effect	Position	Ref	Alt	MAF in tumour	Confirmed by Sanger sequencing
T386-P1282	*MEF2C*	Stopgain	5:88026048	C	A	0.16	No
	MKL1	Non-synonymous	22:40815208	C	A	0.16	No
	PPM1D	**Stopgain**	**17:58740836**	**C**	**T**	**0.23**	**Yes**
	RBM26	**Frameshift**	**13:79940771**	**-**	**T**	**0.21**	**Yes**
	TP53	**Stopgain**	**17:7579389**	**G**	**A**	**0.31**	**Yes**
T438-P1400	***APC***	**Non-synonymous**	**5:112173713**	**G**	**C**	**0.30**	**Yes**
	APC	**Non-synonymous**	**5:112174228**	**G**	**A**	**0.27**	**Yes**
	APC	**Non-synonymous**	**5:112174347**	**G**	**C**	**0.21**	**Yes**
	ARHGAP21	**Frameshift**	**10:24884912**	**T**	**TA**	**0.19**	**Yes**
	ATAD5	**Frameshift**	**17:29214214**	**AT**	**A**	**0.30**	**Yes**
	CARS	**Non-synonymous**	**11:3059285**	**C**	**G**	**0.24**	**Yes**
	CNGB1	**Essential splice site**	**16:57996877**	**C**	**T**	**0.45**	**Yes**
	CORO2B	**Stopgain**	**15:68937546**	**C**	**A**	**0.27**	**Yes**
	FCRL3	**Stopgain**	**1:157667452**	**G**	**A**	**0.23**	**Yes**
	KRT27	**Essential splice site**	**17:38937523**	**C**	**G**	**0.35**	**Yes**
	MLL2	**Non-synonymous**	**12:49448371**	**C**	**T**	**0.26**	**Yes**
	NLRC5	**Frameshift**	**16:57111288**	**TAG**	**T**	**0.32**	**Yes**
	NOTCH2	Frameshift	1:120512275	C	CA	0.17	No
	OR52A5	**Frameshift**	**11:5153238**	**AACCCT**	**A**	**0.17**	**Yes**
	PMS1	**Non-synonymous**	**2:190738302**	**G**	**C**	**0.24**	**Yes**
	SP1	**Stopgain**	**12:53777373**	**C**	**T**	**0.33**	**Yes**
	TP53	**Frameshift**	**17:7578458**	**G**	**GGA**	**0.30**	**Yes**

	TP53	**Non-synonymous**	**17:7578466**	**G**	**A**	**0.36**	**Yes**
	ZNF521	**Non-synonymous**	**18:22804986**	**C**	**G**	**0.19**	**Yes**
	ZNF750	**Frameshift**	**17:80789692**	**G**	**GA**	**0.28**	**Yes**
T437-P1377	*RECQL4*	Non-synonymous	8:145737856	G	T	0.16	No
T441-P1116	***ARHGEF2***	**Frameshift**	**1:155931616**	**TGATAAATACCC**	**T**	**0.16**	**Yes**
	COX6C	**Non-synonymous**	**8:100899805**	**G**	**C**	**0.20**	**Yes**
	FLT3	**Non-synonymous**	**13:28611364**	**C**	**G**	**0.17**	**Yes**
	FZD6	**Stopgain**	**8:104340628**	**C**	**T**	**0.15**	**Yes**
	IL21R	**Non-synonymous**	**16:27459982**	**C**	**T**	**0.23**	**Yes**
	LARGE	**Essential splice site**	**22:33780177**	**C**	**T**	**0.32**	**Yes**
	PAX7	Non-synonymous	1:18961022	G	A	0.16	No
	TET2	**Stopgain**	**4:106156540**	**C**	**T**	**0.32**	**Yes**
	TMPRSS2	Non-synonymous	21:42843880	C	T	0.32	No
T443-P1408	***DNAH10***	**Frameshift**	**12:124285943**	**AG**	**A**	**0.18**	**Yes**
	JUNB	**Frameshift**	**19:12902663**	**C**	**CCT**	**0.33**	**Yes**
	KL	**Frameshift**	**13:33628178**	**AC**	**A**	**0.34**	**Yes**
	OSBPL3	**Stopgain**	**7:24901235**	**T**	**A**	**0.32**	**Yes**
	RORA	**Stopgain**	**15:60789688**	**G**	**T**	**0.16**	**Yes**
	RTL1	**Stopgain**	**14:101348698**	**G**	**A**	**0.17**	**Yes**
	SRPX	**Frameshift**	**X:38080687**	**CAA**	**C**	**0.33**	**Yes**
	VAMP4	**Stopgain**	**1:171678834**	**G**	**A**	**0.21**	**Yes**
T442-P1406	***GPRASP2***	**Stopgain**	**X:101970164**	**A**	**T**	**0.15**	**Yes**
	HPX	**Stopgain**	**11:6458700**	**G**	**A**	**0.18**	**Yes**
	WDR17	**Frameshift**	**4:177083305**	**CT**	**C**	**0.18**	**Yes**
232T-P662	*FAM131C*	Frameshift	1:16384994	C	-	0.10	*
	MYH14	Frameshift	19:50789941	C	-	0.10	No
	TOPAZ1 / C3orf77	Stopgain	3:44283603	C	T	0.10	No
T416-P1354	***CXorf30***	**Frameshift**	**X:36324916**	**AT**	**A**	**0.13**	**Yes**

* Specific primers could not be designed due to similarity with another region

An example of a point mutation and a frameshift mutation are shown in Figure 6.4.

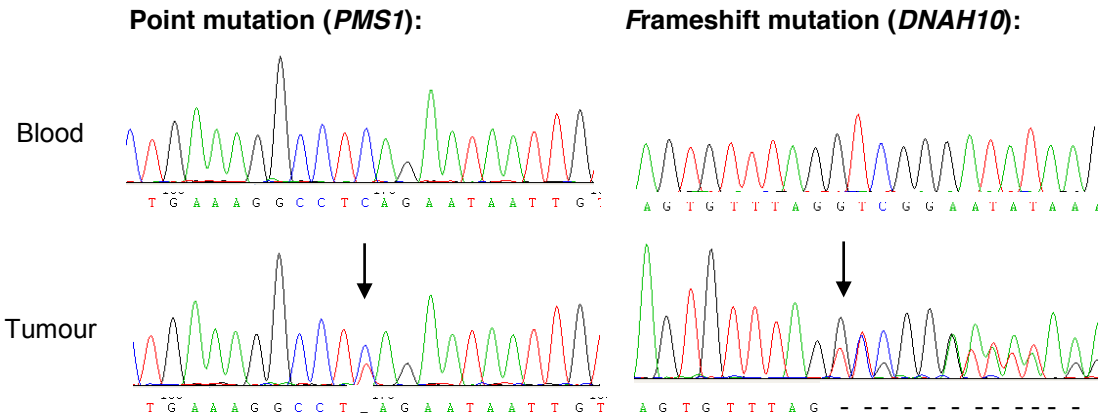


Figure 6.4: Examples of Sanger sequencing chromatograms showing somatic mutations

The mutations that were confirmed are summarized in Table 6.8. Two tumours (T232 and T437) were found not to harbour any somatic mutations that resulted in stop codons, frameshifts or were potential functional variants present in genes in the cancer gene census.

Table 6.8: Somatic mutations confirmed by Sanger sequencing

Sample	Gene	Position	Effect	Protein
386T-P1282	*PPM1D*	17:58740836	Stopgain	R581X
	RBM26	13:79940771	Frameshift	T378fs
	TP53	17:7579389	Stopgain	Q100X
T438-P1400	*APC*	5:112173713	Non-synonymous	D808H
	APC	5:112174228	Non-synonymous	M979I
	APC	5:112174347	Non-synonymous	G1019A
	ARHGAP21	10:24884912	Frameshift	Y20fs
	ATAD5	17:29214214	Frameshift	N1361fs
	CARS	11:3059285	Non-synonymous	V183L
	CNGB1	16:57996877	Essential splice site	-
	CORO2B	15:68937546	Stopgain	Y16X
	FCRL3	1:157667452	Stopgain	Q186X
	KRT27	17:38937523	Essential splice site	-
	MLL2	12:49448371	Non-synonymous	D114N
	NLRC5	16:57111288	Frameshift	I1611
	OR52A5	11:5153238	Frameshift	L210fs
	PMS1	2:190738302	Non-synonymous	E852Q
	SP1	12:53777373	Stopgain	Q541X
	TP53	17:7578458	Frameshift	V157fs
	TP53	17:7578466	Non-synonymous	T155I
	ZNF521	18:22804986	Non-synonymous	E966Q
	ZNF750	17:80789692	Frameshift	P213fs
T441-P1116	*ARHGEF2*	1:155931616	Frameshift	E431fs
	COX6C	8:100899805	Non-synonymous	F52L
	FLT3	13:28611364	Non-synonymous	D423H
	FZD6	8:104340628	Stopgain	R509X
	IL21R	16:27459982	Non-synonymous	T332M
	LARGE	22:33780177	Essential splice site	-
	TET2	4:106156540	Stopgain	Q481X
T443-P1408	*DNAH10*	12:124285943	Frameshift	R742fs
	JUNB	19:12902663	Frameshift	L26fs
	KL	13:33628178	Frameshift	D58fs
	OSBPL3	7:24901235	Stopgain	K311X
	RORA	15:60789688	Stopgain	S513X
	RTL1	14:101348698	Stopgain	R810X
	SRPX	X:38080687	Frameshift	D4fs
	VAMP4	1:171678834	Stopgain	R106X
T442-P1406	*GPRASP2*	X:101970164	Stopgain	K123X
	HPX	11:6458700	Stopgain	R225X
	WDR17	4:177083305	Frameshift	L951fs
T416-P1354	*CXorf30*	X:36324916	Frameshift	L151fs

6.3.4 Function of genes with somatic mutations

The genes found to harbour somatic mutations are shown in Figure 6.9 with a brief description of their function according to the NCBI Gene resource (http://www.ncbi.nlm.nih.gov/gene), unless otherwise stated. Several genes are involved in processes that may contribute to cancer development, including the known tumour suppressors *TP53*, *APC* and *KL*. The last column in Figure 6.9 shows whether the genes may have a potential role in OSCC development, given the known function.

Table 6.9: Function of somatically mutated genes

Tumour	Gene	Full name (from NCBI)	Functional connections	In Cancer Gene Census	Mutation	Potential role in OSCC?
T386-P1282	*PPM1D*	Protein phosphatase, Mg2+/Mn2+ dependent, 1D	Regulates, and is itself regulated by, *TP53*	No	Stopgain	**Yes**
	RBM26	RNA binding motif protein 26	No known function	No	Frameshift	No
	TP53	Tumour protein p53	Cell cycle, apoptosis, DNA repair. A tumour suppressor.	Yes	Stopgain	**Yes**
T438-P1400	*APC*	Adenomatous polyposis coli	Cell migration and adhesion, transcriptional activation, and apoptosis. A tumour suppressor.	Yes	Non-synonymous	**Yes**
	ARHGAP21	Rho GTPase activating protein 21	Cell proliferation and cytoskeleton organization in prostate adenocarcinoma (**a**)	No	Frameshift	**Yes**
	ATAD5	ATPase family, AAA domain containing 5	DNA damage response, apoptosis, DNA replication (**b,c**)	No	Frameshift	**Yes**
	CARS	Cysteinyl-tRNA synthetase	An aminoacyl-tRNA synthetase, located in a tumour-suppressor gene region (11p15.5)	Yes	Non-synonymous	**Yes**
	CNGB1	Cyclic nucleotide gated channel beta 1	Rod photoreceptor	No	Essential splice site	No
	CORO2B	Coronin, actin binding protein, 2B	Reorganization of the neuronal actin cytoskeleton, and potentially neuronal cell migration (**d**)	No	Stopgain	No
	FCRL3	Fc receptor-like 3	Regulation of immune system	No	Stopgain	No
	KRT27	Keratin 27	Forms the cytoskeleton of epithelial cells	No	Essential splice site	No
	MLL2	Myeloid/lymphoid or mixed-lineage leukemia 2	A histone methyltransferase involved in the regulation of several pathways, including p53. (**e**)	Yes	Non-synonymous	**Yes**
	NLRC5	NLR family, CARD domain containing 5	Regulation of immune system	No	Splice site, frameshift	No
	OR52A5	Olfactory receptor, family 52, subfamily A, member 5	Response to smell	No	Frameshift	No
	PMS1	PMS1 postmeiotic segregation increased 1 (S. cerevisiae)	DNA repair	Yes	Non-synonymous	**Yes**
	SP1	Sp1 transcription factor	Cell differentiation, cell growth, apoptosis, immune responses, response to DNA damage, and chromatin remodeling	No	Stopgain	**Yes**
	TP53	As above	As above	Yes	Frameshift, non-synonymous	**Yes**

	ZNF521	Zinc finger protein 521	Regulates B cell development (**f**)	Yes	Non-synonymous	No
	ZNF750	Zinc finger protein 750	Epidermal differentiation (**g**)	No	Frameshift	No
T441-P1116	*ARHGEF2*	Rho/Rac guanine nucleotide exchange factor (GEF) 2	Cell migration (**h**)	No	Frameshift	**Yes**
	COX6C	Cytochrome c oxidase subunit Vic	Catalyzes the electron transfer from reduced cytochrome c to oxygen. Upregulated or involved in gene-fusion events in some tumours (**i**, **j**)	Yes	Non-synonymous	**Yes**
	FLT3	fms-related tyrosine kinase 3	Regulates apoptosis, proliferation, and differentiation of hematopoietic cells in bone marrow	Yes	Non-synonymous	**Yes**
	FZD6	Frizzled family receptor 6	Cell proliferation and apoptosis	No	Stopgain	**Yes**
	IL21R	Interleukin 21 receptor	Proliferation and differentiation of T cells, B cells, and natural killer (NK) cells	Yes	Non-synonymous	**Yes**
	LARGE	Like-glycosyltransferase	Glycosylation of alpha-dystroglycan	No	Essential splice site	No
	TET2	Tet methylcytosine dioxygenase 2	Gene transcription regulation	Yes	Stopgain	**Yes**
T443-P1408	*DNAH10*	Dynein, axonemal, heavy chain 10	Microtubule motor involved in cell division and migration	No	Frameshift	**Yes**
	JUNB	Jun B proto-oncogene	Involved in regulating cell division and tumour invasion (**k**)	No	Frameshift	**Yes**
	KL	Klotho	Tumour suppressor (**l**)	No	Frameshift	**Yes**
	OSBPL3	Oxysterol binding protein-like 3	Intracellular lipid receptor	No	Stopgain	No
	RORA	RAR-related orphan receptor A	Cellular stress response (**m**)	No	Stopgain	**Yes**
	RTL1	Retrotransposon-like 1	Maintenance of the fetal capillaries	No	Stopgain	No
	SPRX	Trypsin-like serine protease	No information	No	Frameshift	No
	VAMP4	Vesicle-associated membrane protein 4	Trafficking synaptic vesicles to the presynaptic membrane	No	Stopgain	No
T442-P1406	*GPRASP2*	G protein-coupled receptor associated sorting protein 2	Regulates G protein-coupled receptors. In HNSCC, gene is overexpressed in patients who develop metastases **(n)**	No	Stopgain	**Yes**
	HPX	Hemopexin	Transports heme	No	Stopgain	No
	WDR17	WD repeat domain 17	Unknown function. Expressed in retina.	No	Frameshift	No
T416-P1354	*CXorf30*	Chromosome X open reading frame 30	No known function	No	Frameshift	No

(**a**) = Lazarini *et al.* 2013; (**b**) = Lee *et al.* 2013; (**c**) = Bell *et al.* 2011; (**d**) = Nakamiyra et al. 1999; (**e**) = Guo *et al.* 2012; (**f**) = Mega *et al.* 2011; (**g**) = Sen *et al.* 2012; (**h**) =Nalbant *et al.* 2009; (**i**) = Kurose *et al.* 2000; (**j**) = Wang *et al.* 1996; (**k**) = Lee and Kim *et al.* 2012; (**l**) =Wang *et al.* 2011; (**m**) = Zhu *et al.* 2006; (**n**) = Rickman *et al.* 2008

6.3.5 Recurrently mutated genes

Genes that are mutated in multiple tumours of the same type are potentially important in the development of OSCC, with the mutations driving the disease. Analysis of protein truncating/splice-site mutations (stopgain, frameshift and essential splice sites) and non-synonymous mutations in genes known to be mutated in cancer (from the Cancer Gene Census) in our data identified only one gene, *TP53*, that was mutated in several tumours. One tumour (T386) contained a *TP53* stop gain mutation, with a second tumour (T438) harbouring both a frameshift insertion and a non-synonymous mutation. If all non-synonymous mutations were included, two additional genes, *GPR98* and *SRRM2,* were found to be mutated in two tumours (and confirmed by Sanger sequencing), as shown in Table 6.10.

Table 6.10: Recurrently mutated genes

Sample	Location (Hg19)	Gene	Protein change	SIFT (score)	Polyphen (score)
T438-P1400	17:578458	*TP53*	GA insertion	-	-
T438-P1400	17:7578466	*TP53*	T155I	Deleterious (0)	Probably damaging (0.938)
T386-P1282	17:7579389	*TP53*	Q100X	-	-
T443-P1408	5:89953731	*GPR98*	I1463T	Tolerated (0.95)	Benign (0.003)
T441-P1116	5:90046431	*GPR98*	R3680C	Deleterious (0)	Probably damaging (1)
T438-P1400	16:2815904	*SRRM2*	G1044E	Tolerated (0.47)	Probably damaging (1)
T441-P1116	16:2809653	*SRRM2*	S275C	Deleterious (0)	Probably damaging (0.999)

The power to detect frequently mutated genes (≥2/8 tumours) is shown in Table 6.11, given different gene mutation rates. For example, if the mutation rate of a gene is 20%, then the probability to detect recurrent mutations (≥2/8 tumours) is 50%.

Table 6.11: Probability of detecting genes recurrently mutated

Gene mutation rate	Probability of ≥2 of 8 tumours being mutated
50%	0.965
40%	0.894
30%	0.745
25%	0.633
20%	0.500
15%	0.343
10%	0.189
5%	0.057

In order to identify more potential driver mutations, the genes which contained confirmed somatic mutations in our study were reviewed for evidence that they were mutated in other published studies in related cancers (Table 6.12). These studies all used a whole-exome sequencing approach to sequence OSCC, oesophageal adenocarcinoma (OAC) and head and neck squamous cell carcinoma (HNSCC). Numbers of tumours in these studies ranged from 11 to 149.

Table 6.12: Comparison of genes mutated in South African OSCC with mutations in related cancers

Sample	Gene mutated in South African OSCC	OSCC Agrawal *et al.* 2012 *n* = 12	OAC Agrawal *et al.* 2012 *n* = 11	OAC* Dulak *et al.* 2013 *n* = 145	HNSCC Agrawal *et al.* 2011 *n* = 32	HNSCC Stransky *et al.* 2011 *n* =74
T386-P1282	*PPM1D*	**Yes**	No	**Yes**	**Yes**	**Yes**
	TP53	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
	RBM26	No	No	**Yes**	No	**Yes**
T438-P1400	*APC*	No	**Yes**	**Yes**	No	**Yes**
	ARHGAP21	No	No	**Yes**	No	No
	ATAD5	No	No	**Yes**	No	No
	CARS	No	No	**Yes**	No	No
	CNGB1	No	No	**Yes**	No	**Yes**
	CORO2B	No	No	**Yes**	No	**Yes**
	FCRL3	No	No	**Yes**	No	**Yes**
	KRT27	No	No	No	No	No
	MLL2	**Yes**	No	**Yes**	No	**Yes**
	NLRC5	No	No	No	No	**Yes**
	OR52A5	No	No	**Yes**	No	**No**
	PMS1	No	No	**Yes**	No	**Yes**
	SP1	No	No	**Yes**	No	No
	TP53	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
	ZNF521	No	No	**Yes**	No	**Yes**
	ZNF750	**Yes**	No	**Yes**	No	**Yes**
T441-P1116	*ARHGEF2*	No	No	**Yes**	No	**Yes**
	COX6C	No	No	No	No	No
	FLT3	No	No	**Yes**	No	**Yes**
	FZD6	No	No	**Yes**	No	No
	IL21R	No	No	**Yes**	No	No
	LARGE	No	No	No	No	**Yes**
	TET2	No	No	No	No	**Yes**
T443-P1408	*DNAH10*	No	No	**Yes**	No	**Yes**
	JUNB	No	No	**Yes**	No	No
	KL	No	No	**Yes**	No	No
	OSBPL3	No	No	**Yes**	No	No
	RORA	No	No	**Yes**	No	**Yes**
	RTL1	No	No	**Yes**	No	No
	SPRX	No	No	No	No	No
	VAMP4	No	No	No	No	No
T442-P1406	*GPRASP2*	**Yes**	No	**Yes**	No	**Yes**
	HPX	No	No	No	No	**Yes**
	WDR17	No	No	**Yes**	No	**Yes**
T416-P1354	*CXorf30*	No	No	No	No	No

* Data only available for genes that were mutated in >1 tumour

Five of the 6 blood-tumour pairs (excluding the 2 tumours where no mutations were identified) contained mutations in genes that are mutated in either OSCC, OAC or HNSCC, thereby providing support that driver mutations are present in the South African OSCC samples. Interestingly, 28 of the 37 mutated genes (75.7%) were also mutated in more than one OAC tumour in the study by Dulak *et al.* (2013). This study whole-exome sequenced 149 patients, showing that a large number of samples are needed to identify recurrent mutations. In the OSCC study by Agrawal *et al.* (2012), mutations were present in 5 of the 37 (13.5%) genes that were mutated in South African OSCC patients. This lower number is probably due to the relatively small number of samples sequenced in the Agrawal *et al.* study (*n*=12).

GPR98 and *SRRM2* are also mutated in other OSCC, OAC and HNSCC exome sequencing studies, as shown in Table 6.13.

Table 6.13: *GPR98* and *SRRM2* somatic mutations in published studies

Gene mutated in South African OSCC	OSCC	OAC	OAC*	HNSCC	HNSCC
	Agrawal *et al.* 2012	Agrawal *et al.* 2012	Dulak *et al.* 2013	Agrawal *et al.* 2011	Stransky *et al.* 2011
	n = 12	*n* = 11	*n* = 145	*n* = 32	*n* =74
GPR98	Yes	No	Yes	No	Yes
SRRM2	Yes	No	Yes	No	Yes

* Data only available for genes that were mutated in >1 tumour

6.3.6 *TP53* sequencing

As *TP53* is the most commonly mutated gene in cancer (Efeyan and Serrano 2007), this gene was further investigated by Sanger sequencing the coding regions of the gene in all of the 10 blood-tumour pairs that were available. Eight of these had been whole-exome sequenced, but Sanger sequencing could potentially identify mutations that were missed. In total, 8 somatic mutations were identified in 6 tumours, giving a *TP53* mutation rate of 60%. The mutations are summarized in Table 6.14, with the sequencing chromatograms shown in Figure 6.5. Of these eight mutations, two were frameshifts, four were non-synonymous, one was a stop mutation and one was in a splice site.

Table 6.14: *TP53* somatic mutations identified by Sanger sequencing

Sample	Exon	Position - Chr 17 (build 37)	SNP ID	Mutation ID (a)	Location	Genotypes		Triplet code (blood --> tumour)	mRNA	Protein
						Blood	Tumour			
T386-P1282	4	7579389	-	COSM44032	Stop	C/C	C/T	CAG --> TAG	298C>T	Gln100X
T443-P1408	4	7579358	rs11540654	COSM10716	Non-synonymous	G/G	G/T	CGT --> CTT	329G>T	Arg110Leu
T438-P1400	5	7578466	-	COSM44033	Non-synonymous	C/C	C/T	ACC --> ATC	464 C>T	Thr155Ile
T438-P1400	5	7578458	-	-	Frameshift	WT	TC ins / WT		471-472	Val157fs
GSH-P1508	5	7578407	rs138729528	COSM10870	Non-synonymous	C/C	G/C	CGC --> GGC	523C>G	Arg175Gly
T441-P1116	6	7578177	-	COSM44014	Splice site/ synonymous	G/G	G/A	GAG -- GAA	672G>A	Glu224Glu
T441-P1116	7	7577567 - 7577572 (**b**)	-	-	Frameshift	WT	TG ins / WT		~709-714 (b)	M237fs or C238fs (b)
T442-P1406	8	7577100	-	COSM11123	Non-synonymous	A/A	A/G	AGA --> GGA	838A>G	Arg280Gly

(a) Mutation ID from COSMIC database (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/)
(b) Exact position could not be determined due to poor sequence quality in the forward direction

mRNA: NM_001126112.1; Protein: NP_001119584.1

221

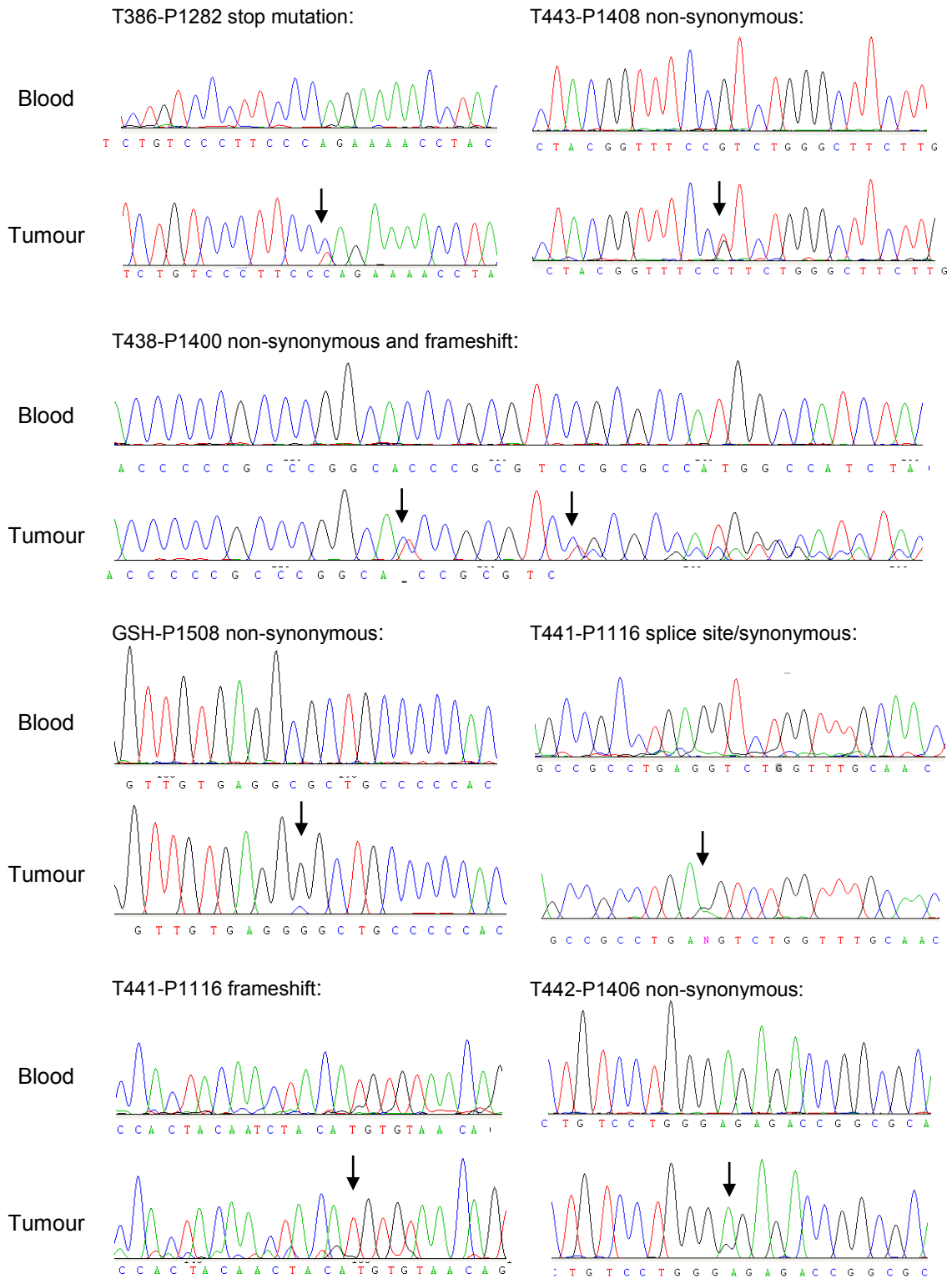


Figure 6.5: Chromatograms for *TP53* mutations identified by Sanger sequencing

Three of the four non-synonymous mutations are predicted to be probably damaging by Polyphen and deleterious by SIFT (Table 6.15). The fourth mutation, Arg110Leu, is predicted to be possibly damaging and tolerated by each program, respectively, but this does vary depending on the *TP53* transcript. For example, in the ENST00000503591 transcript the mutation is predicted to be deleterious in SIFT (score = 0.02).

Table 6.15: Functional predictions of *TP53* non-synonymous mutations
Based on *TP53* ENST00000269305 transcript.

Sample	Mutation	Polyphen (score)	SIFT (score)
T443-P1408	Arg110Leu	Possibly damaging (0.46) *	Tolerated (0.06) *
T438-P1400	Thr155Ile	Probably damaging (0.938)	Deleterious (0)
GSH-P1508	Arg175Gly	Probably damaging (0.98)	Deleterious (0)
T442-P1406	Arg280Gly	Probably damaging (0.982)	Deleterious (0)

* Prediction varies depending on *TP53* transcript

Seven of the eight somatic mutations were located in tumours that had been whole-exome sequenced. Of these, 4 had been detected by this method but 3 mutations were not, which may be due to the sequence reads not meeting the thresholds to call a mutation. The exome sequencing data was reviewed in IGV for these mutations, with screen shots of these mutations shown below.

Figure 6.6 shows the Arg110Leu mutation present in T443-P1408, where leucine (A allele) was present in 34% of reads in the tumour and absent in the blood. The total number of reads at this position was 90 and 80 for blood and tumour DNA, respectively. The mutation was not identified by the ICR exome sequencing analysis pipeline as it was mis-identified as a common SNP and removed from the analysis.

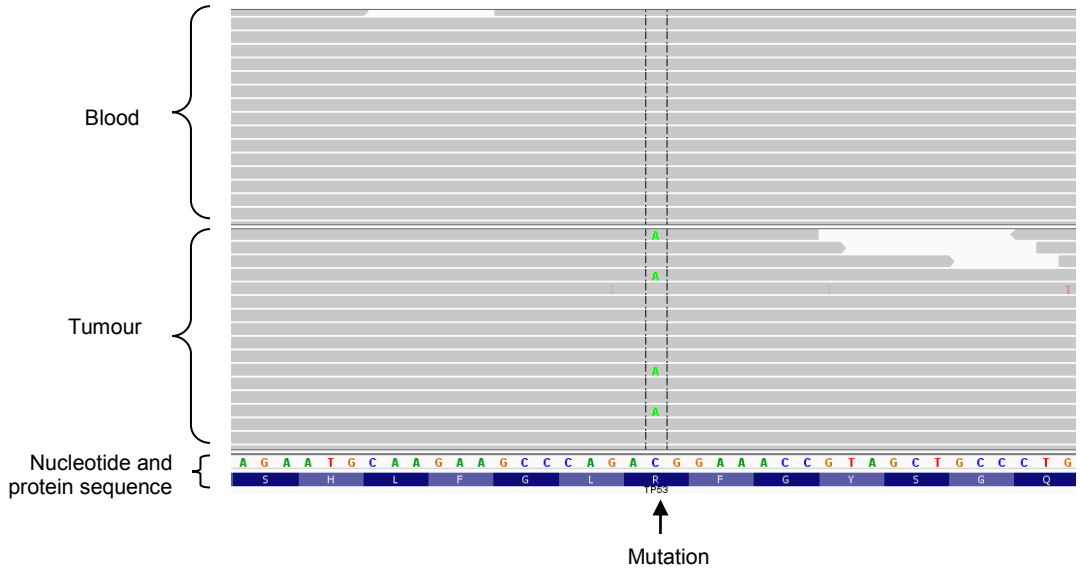


Figure 6.6: Exome sequencing reads of *TP53* Arg110Leu mutation
The somatic mutation was identified in the blood-tumour pair T443-P1408. Exome sequencing data was visualised on IGV.

Figure 6.7 shows the Arg280Gly mutation present in T442-P1406, where glycine (C allele) was present in 12% of reads in the tumour (below the 15% threshold) and absent in the blood The total number of reads at this position was 43 and 34 for blood and tumour DNA, respectively.

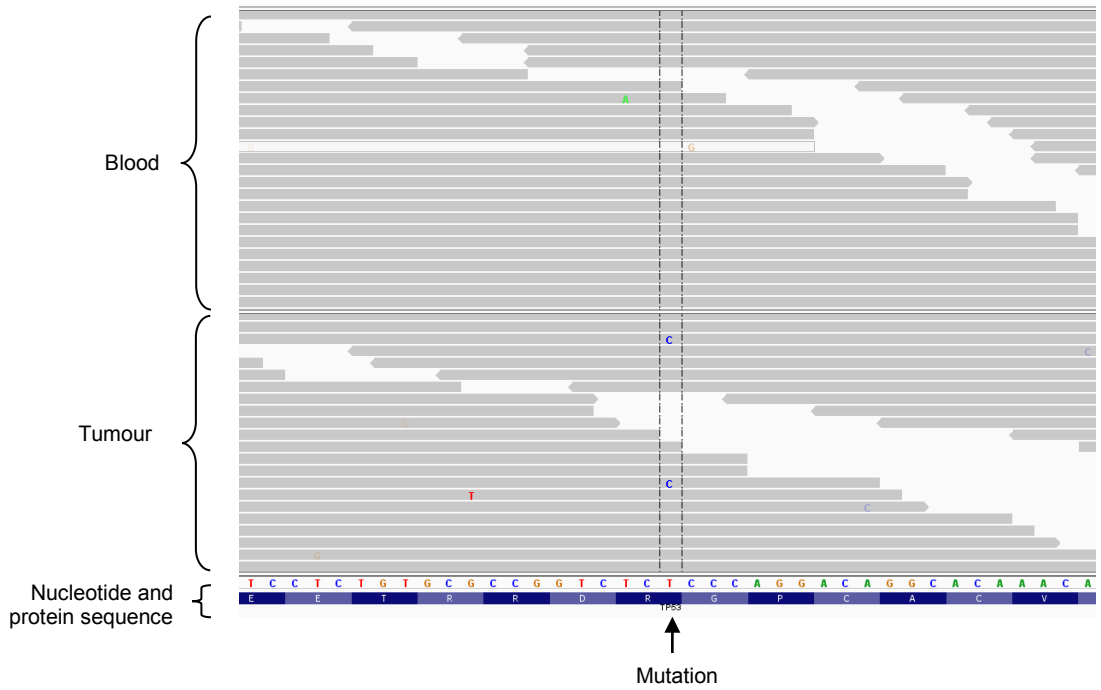


Figure 6.7: Exome sequencing reads of *TP53* Arg280Gly mutation
The somatic mutation was identified in the blood-tumour pair T442-P1406. Exome sequencing data was visualised on IGV.

Figure 6.8 shows the frameshift insertion in exon 7 present in T441-P1116, where the insertion was present in 6% of reads in the tumour (below the 15% threshold) and absent in the blood. The total number of reads at this position was 29 and 50 for blood and tumour DNA, respectively.

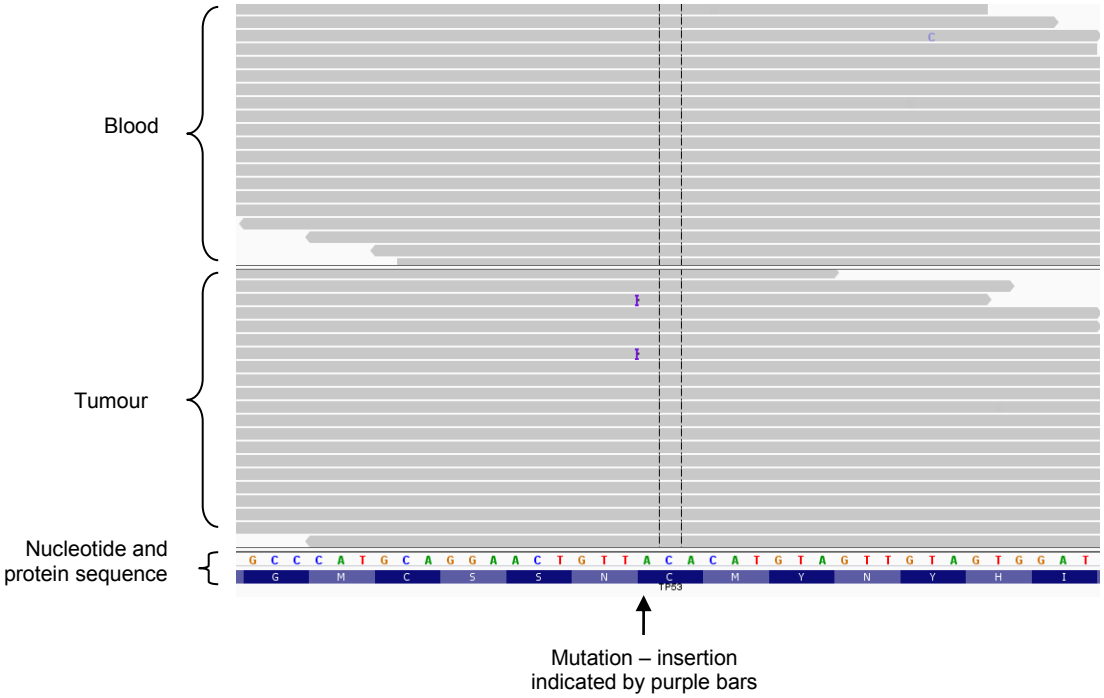


Figure 6.8: Exome sequencing reads of *TP53* exon 7 frameshift insertion
The insertion was identified in the blood-tumour pair T441-P1116. Exome sequencing data was visualised on IGV.

6.3.7 *PPM1D* sequencing

A stopgain mutation was identified in *PPM1D* by whole-exome sequencing. As this gene is involved in the *TP53* pathway, the exons of *PPM1D* were Sanger sequenced in all available 11 blood-tumour pairs. Sequencing was incomplete for one tumour (288T) which had a low yield. In total, six somatic mutations were identified: one in the 5' UTR, two synonymous, two identical non-synonymous mutations, and one resulting in a stop-codon. These were present in four of the blood-tumour pairs, with seven tumours not containing a somatic mutation in *PPM1D.* The results are summarized in Table 6.16, with the chromatograms shown in Figure 6.9.

Table 6.16: *PPM1D* somatic mutations identified by Sanger sequencing

Tumour-blood	Exon	Position – Chr 17 (build 37)	SNP ID	Location	Genotypes Blood	Genotypes Tumour	Mutation/ variant	mRNA	Protein
T386-P1282	1	58677580	rs116268471	5' UTR	T / C	T / -	-	-	-
T386-P1282	1	58677865	rs16944543	Synonymous	G / A	G / -	GAG --> GAA	90G>A	Glu30Glu
T386-P1282	6	58740836	Novel	Stop mutation	C / C	C / T	CGA --> TGA	1741 C>T	Arg581X (tumour)
288T-P920	5	58734091	rs111239559	Synonymous	A / C	A / -	GGA --> GGC	1149 A>C	Gly383Gly
GSHT-P1508	6	58740785	Novel	Non-synonymous	G / C	G / -	GCA --> CCA	1690 G>C	Ala564Pro (blood)
TBHT-TB62	6	58740785	Novel	Non-synonymous	G / C	G / -	GCA --> CCA	1690 G>C	Ala564Pro (blood)

mRNA: NM_003620.3; Protein: NP_003611.1

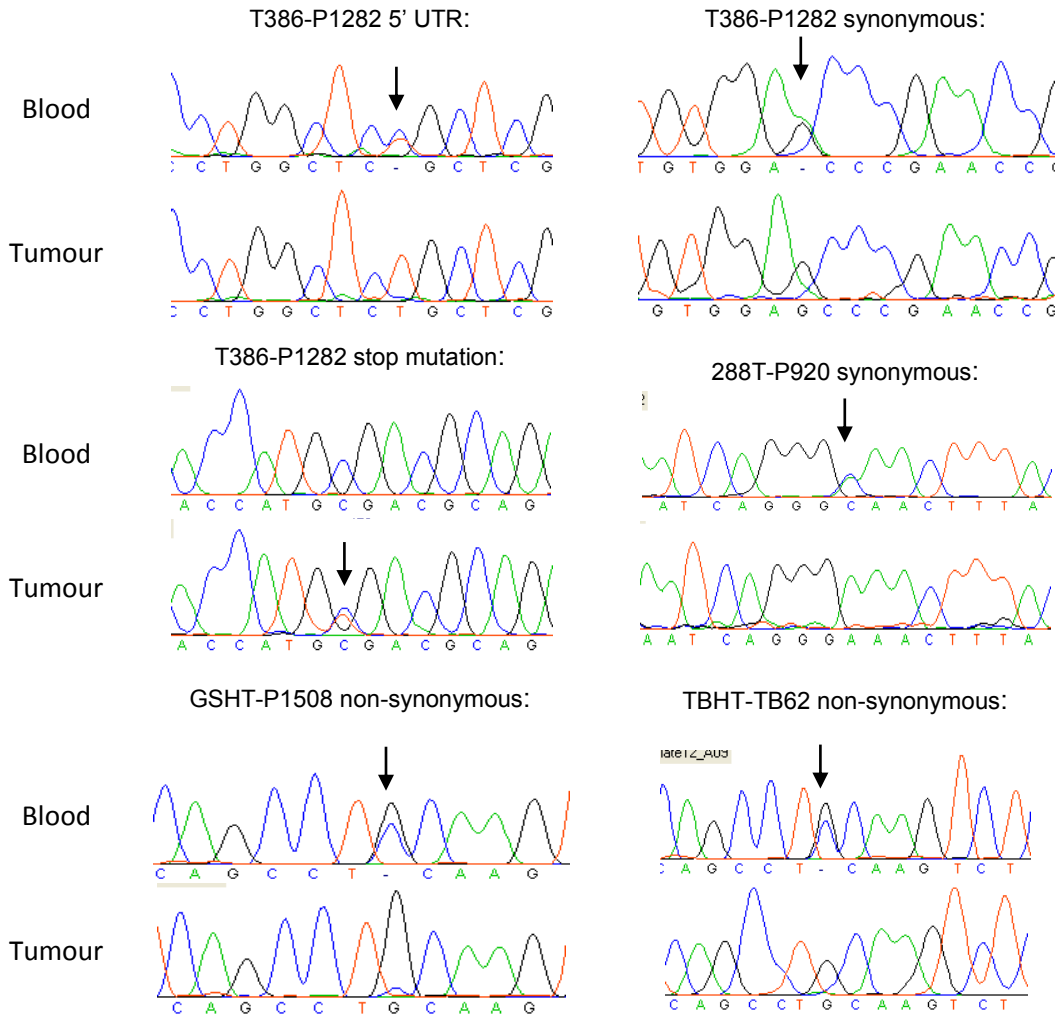


Figure 6.9: Sanger sequencing chromatograms of *PPM1D* somatic mutations

The stop codon (Arg581X) is a novel mutation located in exon 6 of *PPM1D*. The germline DNA was homozygous for the wildtype allele (Arg/Arg) whilst the tumour was heterozygous (Arg/X). The non-synonymous mutation found in both GSHT-P1508 and TBHT-TB62 was a novel variant resulting in a alanine to proline substitution at position 564. The variant was heterozygous (Ala/Pro) in the blood, but showed loss of heterozygosity (LOH) in the tumour (Ala/Ala or Ala/-). The change from alanine to proline is predicted to be possibly damaging by Polyphen but tolerated by SIFT. The two synonymous mutations, Glu30Glu and Gly383Gly, and the 5' UTR mutations are all known variants, and all are heterozygous in the blood and show LOH in the tumour. A total of 4 of 11

tumours showed LOH at the *PPM1D* locus, and one of these also harboured a stop mutation.

6.4 Discussion

Whole-exome sequencing was performed in 8 matching blood-tumour pairs from South African OSCC patients. The median depth of sequencing reads ranged from 52 to 283, with >84% of targeted regions covered at >20X and >63% covered at >40X. The samples sequenced by Illumina achieved the highest coverage (median depth of >244) due to only one sample being sequenced per lane of the flow cell, compared to multiple samples for those sequenced at KCL and ICR.

The number of potential somatic mutations in the tumours ranged between 10 and 301, with an average of 117 per tumour. The thresholds which were used to identify a somatic mutation are discussed below.

6.4.1 Thresholds used in somatic mutation identification

The initial criteria for calling a somatic mutation was based on the variant being essentially absent in the blood (<2% to allow for sequencing errors) and present in ≥15% of sequencing reads in the tumour. However, in two blood-tumour pairs (232T-P662 and T416-P1354), a low number of somatic mutations were identified in each tumour using this threshold (6 and 3 mutations, respectively). This lack of mutations was likely the result of a high level of normal tissue contamination in the tumour sample. An alternative explanation could be that only a small number of mutations are needed to cause cancer development. However, in one blood-tumour pair (T416-P1354), no protein truncating/splice-site mutations (stopgain, frameshift or essential splice site) somatic mutations were identified, suggesting that no driver mutations were present. The other blood-tumour pair (232T-P662) only contained one protein altering mutation. As a result, the threshold for the percentage of sequencing reads supporting the alternative allele in the tumour was lowered to ≥10% for these two samples, to compensate for a higher level of normal tissue contamination.

Decreasing the threshold resulted in a modest increase in potential somatic mutations, with 55 and 10 mutations identified in the two tumour pairs (232T-P662 and T416-P1354, respectively), which is at the lower end of the number of somatic mutations detected in all tumours. The number of protein truncating/splice-site mutations present increased to 10 and 1 for the tumour pairs, respectively. It may be very difficult to identify valid somatic mutations in tumour T416 particularly, as false-positives would become more common if the threshold was lowered any further. Alternatively, there may in fact be very few point mutations in these particular tumours (see section 6.4.4)

There is no consensus for what the threshold in the tumour should be. Several other cancer studies have used ≥15% (see Table 6.17), which allows for ~33% of the tissue sample to be derived from tumour DNA. Other published studies use a ≥10% threshold or do not state a value. A recent study has used varying thresholds depending on cross-contamination estimates for each sample (Dulak *et al.* 2013).

Table 6.17: Thresholds used for somatic mutation calling in published exome sequencing studies

Study	% reads containing mutation in tumour	Minimum total reads required in:	
		Blood	Tumour
Agrawal *et al.* 2012	15%	-	-
Dulak *et al.* 2013	Per-sample basis	8	14
Stransky *et al.* 2011	-	8	14
Agrawal *et al.* 2011	10% or 15%	9	-
Le Gallo *et al.* 2012	-	5	5
Liu *et al.* 2012	15%	8	15

"-" = No information available

Another threshold imposed stipulates the number of sequencing reads that must be present. In this analysis, ≥8 and ≥14 total reads in the blood and tumour, respectively, were required to call a somatic mutation. This threshold was chosen based on published studies, for example, those by Dulak *et al.* (2013) and Stransky *et al.* (2011), see Table 6.17.

6.4.2 Confirmation of somatic mutations

The first step to confirm potential mutations identified by exome sequencing was to visualise the sequencing data using the Integrative Genomics Viewer (IGV). Mutations were confirmed to be accurately called as somatic mutations if they were absent in the normal DNA and clearly present in the tumour. However, some variants did not appear to be somatic mutations, mainly as they were located in a region with many other variants, both in blood and tumour, suggesting that it might represent mis-alignment or sequencing errors. Alternatively, some variants were also present in the normal DNA. Visualising all protein truncating/splice-site mutations and non-synonymous mutations in genes in the Cancer Gene Census produced varying results, with some tumours having a large proportion of mutations called correctly (100%), and others with a high degree of false-positives (9% confirmed). The rate of mutation confirmation using IGV was much higher for the exomes analysed by the ICR than those at KCL (95-100% vs. 9-30%, respectively). However, using the KCL calling algorithm on the samples that were originally called by the ICR, a similar number of mutations were confirmed by IGV, although there were more false positives, suggesting the specificity of the KCL method is lower but the sensitivity is higher compared to the ICR pipeline. Therefore, the low percentage confirmed for the 3 samples only analyzed by KCL is unlikely to be due to a sub-optimal mutation calling pipeline.

The large number of synonymous and intronic mutations detected were not analysed by IGV due to time constraints and the low probability that they represent driver mutations in these tumours.

The mutations that appeared to be valid using IGV were then analysed by Sanger sequencing. In four of the tumours, all somatic mutations were confirmed. In the other four tumours, confirmation rates were 95% (20/21 mutations), 30% (3/10), 24% (5/21) and 9% (1/11). The reasons for the lack of somatic mutations in some tumours are discussed in 6.4.4.

6.4.3 Genes with somatic mutations

Protein truncating/splice-site mutations were identified in several genes that have functions including regulation of the cell cycle and apoptosis. This suggests that, if mutated, they may be involved in tumourigenesis, and hence, are plausible driver mutations. For example, stop mutations were identified in *SP1*, *PPM1D* and *FZD6,* which are all involved in cell proliferation and growth, and in apoptosis. However, whether these mutations actually are driver mutations is difficult to determine from our data as, apart from *TP53*, none of the genes were found to be recurrently mutated, as discussed below.

6.4.3.1 Recurrently mutated genes

For mutations likely to cause protein truncation, are in essential splice-sites or are non-synonymous mutations in genes that are present in the Cancer Gene Census, only one gene contained mutations in more than one tumour. This gene, *TP53*, was mutated in 2/8 tumours (25%) based on the exome sequencing data. When all non-synonymous mutations are considered, not just those in the Cancer Gene Census, two additional genes are found to be recurrently mutated in the South African OSCC patients. Both *GPR98* and *SRRM2* contain non-synonymous mutations in two tumours each (25% mutation rate), which were confirmed by Sanger sequencing.

GPR98 (G protein-coupled receptor 98) has been found to be involved in several processes and diseases, including the regulation of bone mineral density (Urano *et al.* 2012) and Usher syndrome which causes deafness and blindness (Weston *et al.* 2004; Ebermann *et al.* 2009). The role of *GPR98* in cancer development is not clear, although it is frequently mutated in melanoma (27.5% of tumours) (Prickett *et al.* 2011), and is mutated in Barrett's oesophagus (Streppel *et al.* 2013). G protein-coupled receptors are known to regulate functions including cell proliferation and survival (reviewed in Dorsam

and Gutkind 2007), and so mutations in *GPR98* may be shown to alter key pathways leading to cancer formation.

SRRM2 (serine/arginine repetitive matrix 2) is an RNA splicing factor that has been shown to be differentially expressed in Parkinson's disease (Shehadeh *et al.* 2010). The RNA binding protein is also involved in cell migration in ovarian cancer cell lines (Mukherji *et al.* 2006), and is deregulated in colorectal cancer patients (Wu *et al.* 2012.b).

GPR98 and *SRRM2* are both mutated in other OSCC, OAC and HNSCC exome sequencing studies (Stransky *et al.* 2011; Agrawal *et al.* 2012; Dulak *et al.* 2013). However, although mutated in up to 12% of samples (9/74) (Stransky *et al.* 2011), neither gene is reported to be significantly mutated in the studies by Dulak *et al.* (2013) and Stransky *et al.* (2011). Significant mutations were identified by comparing the observed number of mutations in each gene to the number expected by chance given the background mutation rate. The non-significant mutations are suggested more likely to be passenger mutations, rather than driver mutations (Stransky *et al.* 2011).

Only 3 genes (*TP53*, *GPR98* and *SRRM2*) were recurrently mutated in the South African OSCC tumours, and the significance of the *GPR98* and *SRRM2* mutations is not yet clear. Our pilot study lacked power to detect other relatively frequently mutated genes, with only a 50% probability to detect recurrent mutations (≥2/8 tumours) in a gene with a 20% mutation rate. Therefore, more tumours need to be sequenced in order to identify genes that contain driver mutations in OSCC in South African populations. Promisingly, of the 37 genes that were mutated, 28 (75.7%) were also recurrently mutated in OAC (Dulak *et al.* 2013), providing evidence that driver mutations may be present.

6.4.3.2 *TP53*

p53 is a transcription factor that is activated in response to cellular stress, such as DNA damage (reviewed in Vousden and Lu 2002). Activation of p53 enables cells to undergo cell-cycle arrest, thereby preventing DNA replication occurring in these abnormal conditions, following which they may re-enter the cell cycle or undergo apoptosis.

TP53 is one of the most commonly mutated genes in human cancers, with approximately 50% of tumours harbouring a mutation (Vousden and Lu 2002). In OSCC, mutation rates range from 17 to 84% (reviewed in Egashira *et al.* 2007). In the South African OSCC patients described in this chapter, six out of 10 tumours (60%) contained somatic mutations in *TP53*. Of the 8 mutations identified by Sanger sequencing, one was a stop-mutation, two were frameshift insertions, 4 were non-synonymous, and 1 was at a splice site (the last base-pair of the exon). All of the mutations were located in the region encoding the DNA binding domain (residues 92-292), where the majority (97%) of missense mutations have previously been shown to reside (Olivier *et al.* 2002). The stop mutation and frameshifts are likely to alter the function of the protein. The four non-synonymous mutations are also all predicted to be deleterious and probably damaging by SIFT and Polyphen, respectively, in one or more *TP53* gene transcripts, thereby also possibly affecting protein function.

The exome-sequencing data called only 4 of the 7 *TP53* mutations that were identified by Sanger sequencing (1 additional mutation was in a tumour sample that was not exome sequenced). Two mutations did not reach the required threshold for the percentage of reads supporting the mutation in the tumour sample (≥15%). These were Arg280Gly in T442-P1406 (12% of reads) and a frameshift insertion in T441-P1116 (6% of reads). A third mutation, Arg110Leu, in T443-P1408 was present in 34% of reads in the tumour and absent in the blood according to IGV, with a high number of sequencing reads at the position (>80). The variant was not identified as a somatic mutation by the ICR analysis

pipeline as it was mis-identified as a common SNP and removed from the analysis.

6.4.3.3 *PPM1D*

A stop-gain somatic mutation in *PPM1D* was identified in the blood-tumour pair T386-P1282. *PPM1D* (protein phosphatase, Mg2+/Mn2+ dependent, 1D) is a proto-oncogene involved in DNA damage response (Bulavin *et al.* 2002). PPM1D is involved in a negative feedback loop, both regulating the activity of p53 and itself being regulated by p53 *(reviewed in Lu et al. 2005)*. Briefly, following DNA damage, p53 is activated through its phosphorylation by a number of proteins to result in cell-cycle arrest leading to the repair of DNA damage or apoptosis. *TP53* activation also leads to an increased expression of *PPM1D*, which then causes the dephosphorylation of p53 either directly or indirectly, leading to decreased p53 activity (reviewed in Lu *et al.* 2008). This allows the cell to return to the cell-cycle after DNA repair.

Changes in *PPM1D* expression may have knock-on effects on *TP53* which could result in mis-regulation of the tumour suppressor gene. For example, overexpression of *PPM1D* may prevent the activation of *TP53* resulting in cells that are unable to undergo apoptosis (Lu *et al.* 2008). Indeed, *PPM1D* mRNA overexpression has been observed in a number of cancers including breast, neuroblastomas, ovarian clear cell carcinomas and gastric cancer (Bulavin *et al.* 2002; Li *et al.* 2002; Saito-Ohara *et al.* 2003; Fuku *et al.* 2007; Tan *et al.* 2009). In addition, *PPM1D* inactivation is shown to suppress tumourigenesis (Bulavin *et al.* 2004). Somatic mutations in *PPM1D* have also been identified including in head and neck squamous cell carcinoma (HNSCC) and OSCC (Agrawal *et al.* 2011; Stransky *et al.* 2011; Agrawal *et al.* 2012). However, the mutations are not recurrent in each study (1/12 OSCC patients and 1/32 HNSCC cases), and hence, are not reported in the main text of published papers (Agrawal *et al.* 2011; Agrawal *et al.* 2012). In addition to somatic mutations, mosaic germline

mutations have recently been identified in *PPM1D* which predispose to breast and ovarian cancer (Ruark *et al.* 2013).

This evidence suggests that *PPM1D* is an important gene for cancer development, and hence, the stop-gain mutation (Arg581X) identified through whole-exome sequencing may play a role in OSCC tumourigenesis. This mutation was absent in the blood (P1282) and was heterozygous in the matching tumour (T386). The mutation, located in exon 6, is not located in a known protein domain and is downstream of the catalytic phosphatase domain and the nuclear localization signal, suggesting that the protein may still be functional. This is consistent with functional studies by Ruark *et al.* (2013) which show that truncating mutations in this C-terminal region result in enhanced suppression of p53 in response to ionising radiation, suggesting that the mutant alleles encode hyperactive PPM1D isoforms.

The seven exons of *PPM1D* were Sanger sequenced in all available blood-tumour pairs to further investigate somatic mutations. Although 8 of the 11 matched pairs had been whole-exome sequenced, they were Sanger sequenced to ensure that no mutations had been missed. The stop mutation was confirmed, but no other somatic mutations were detected. However, four patients (36%) were heterozygous for one or more germline variants in *PPM1D*, with loss of heterozygosity occurring in their respective tumours. Only one of these was a non-synonymous mutation, with two synonymous mutations and a mutation in the 5' UTR also identified. This data suggests that LOH at chromosome 17q23.2 may be common in these tumours. It would be important to determine the extent of the loss to establish whether other genes might be involved. One tumour showing LOH (T386) was exome sequenced, allowing neighbouring regions to be assessed for LOH. Using IGV to visualise the data, no LOH was observed in a ~2,000 kb region surrounding *PPM1D*. Other studies have analyzed LOH and copy number alterations in OSCC, with several

identifying LOH at chromosome 17q (Hu *et al.* 2006; Hu *et al.* 2009; Chattopadhyay *et al.* 2010), although none included the 17q23.2 *PPM1D* locus.

The non-synonymous mutation, Ala564Pro, is located in exon 6. The variant was identified in two patients (GSHT-P1508 and TBHT-TB62), and was heterozygous in the blood but absent in the tumour, indicative of LOH. The variant is novel and is predicted to be possibly damaging by Polyphen but tolerated by SIFT. It is located just downstream of the cluster of germline truncating mutations in exon 6 of *PPM1D* (amino acids 420-546) previously associated with an increased risk of breast and ovarian cancers (Ruark *et al.* 2013), and it is therefore possible that the Ala564Pro mutation in the germline DNA may represent a susceptibility variant for OSCC. This would need to be determined using a large case-control association study, and the potential functional effect of the mutation investigated.

6.4.4 Samples which lack somatic mutations

Two tumours did not contain any somatic mutations that were likely to be driver mutations (either protein truncating/splice-site mutations or non-synonymous mutations present in the Cancer Gene Census). T437 and T232 contained 1 and 3 potential mutations, respectively, but none were confirmed by Sanger sequencing. The threshold for the percentage of alternative alleles in the tumour was set at ≥15% for T437, as 171 somatic mutations were initially identified by exome sequencing. Therefore, lowering this threshold to ≥10% may identify true somatic mutations that are possibly drivers. The tumour T232 used the lower threshold of ≥10% sequencing reads supporting the alternative allele. This threshold allows the detection of heterozygous mutations in 'tumour DNA' where only ~33% of the sample is derived from tumour tissue. If this were reduced further, false positives may be introduced.

The most likely explanation for the lack of potential driver mutations is that there was a high degree of normal tissue contamination within the tumour samples.

The percentage of tumour tissue present in each sample was not known since it is the practice of our collaborators to take multiple biopsies from patients, and the biopsy used for histological diagnosis of OSCC is not the one used for DNA extraction. Any further exome sequencing studies should ensure that an estimate of the tumour tissue content is provided through histological examination of tissue sections taken from the same biopsy used for DNA extraction, and that those with the highest estimate are selected for sequencing. Normal tissue could also be removed from specimens using macrodissection (Biankin *et al.* 2012). However, most patients with OSCC are not resected, so limiting amounts of tissue are available from biopsies.

Alternatively, an estimate of the amount of tumour tissue in a sample could be obtained by genotyping matching blood and tumour DNA on a SNP genotyping array (Song *et al.* 2012). If the tumour contains regions of LOH, the amount of contamination of the sample with normal tissue can be quantified from the shift in SNP allele frequency at regions of LOH. Software (qpure) to implement this method is available (Song *et al.* 2012). This approach has been used in a study of pancreatic cancers, where samples successfully whole-exome sequenced contained as little as 20% tumour DNA (Biankin *et al.* 2012).

6.4.5 Summary

The exomes of 8 OSCC blood-tumour pairs were sequenced, with five tumours successfully sequenced using 500 ng of starting DNA, instead of the recommended 3 µg. The total number of potential somatic mutations varied between 10 and 301, although some of these are likely to be artefacts, as shown by visualisation of the data using IGV. Two samples did not appear to contain any potential driver mutations, which was most likely due to a high contamination with normal tissue DNA. In the other 6 tumours, mutations were identified in known tumour suppressor genes including *TP53*, *KL* and *APC*. Recurrent mutations only occurred in 3 genes, *TP53*, *GPR98* and *SRRM2*, which is probably due to the relatively low number of samples that were

sequenced. The latter two genes contained only non-synonymous mutations which may not alter protein function. However, it is promising that the majority of genes that were mutated have previously been mutated in either OSCC, oesophageal adenocarcinoma or head and neck squamous cell carcinoma. Sequencing additional OSCC samples may enable these mutations to be confirmed as recurrent, providing evidence that they may be driver mutations for the disease.

7 Discussion

7.1 Key findings

7.1.1 Genetic susceptibility to OSCC

This thesis aimed to identify genetic variants associated with OSCC in the South African Black and Mixed Ancestry populations. Initially, candidate gene association studies were performed. In Chapter 3, variants were selected based on their evidence of association with OSCC in candidate gene studies in other populations, namely Asian and European populations. Thirteen variants were tested for association with the disease in the South African populations, and only one, *ALDH2* +82 A>G (rs886205) was significantly associated with OSCC (P=0.0038). In addition, a further four variants had a suggestive association with the disease (P<0.05). All of these were identified in the Mixed Ancestry population, with no variants showing evidence of association in the Black population.

Following this work, three independent OSCC GWAS were published in the Chinese population (Abnet *et al.* 2010; Wang *et al.* 2010.a; Wu *et al.* 2011.c). These identified a total of 8 SNPs in 6 loci associated with the disease, including *PLCE1* His1927Arg (rs2274223), which was associated in all studies. In Chapter 4 of this thesis, these variants were investigated in the South African populations using a case-control association study. Only *RUNX1* rs2014300 was associated with OSCC in the Mixed Ancestry population, with no variants associated in the Black population. Due to the strong evidence of the involvement of *PLCE1* in OSCC development in the Chinese population, this gene was further investigated in the South African Black population by sequencing all of the exons in 46 individuals to identify potential functional variants present in this population. This led to five variants being selected for follow-up genotyping, based on amino acid conservation across species and whether the mutations were predicted to be damaging. In a case-control association study, one of these variants, *PLCE1* Arg548Leu (rs17417407), was

significantly associated with OSCC in the Black population, with a minor allele frequency of 16.6% in cases and 21.1% in controls (P = 0.008). This was the first variant in any of our studies to be significantly (or suggestively) associated with OSCC in the Black population. Although this variant was not genotyped in the Chinese GWAS studies, a SNP in high LD (r^2 = 1) with it, rs2689700, was not reported as being associated with OSCC. *PLCE1* Arg548Leu was not in LD with His1927Arg in either the Chinese/Japanese HapMap population or the South African Black populations, suggesting that Arg548Leu is an independent risk variant.

The genetic susceptibility to OSCC was further investigated using the Immunochip, a custom genotyping platform designed for immune related disorders. This array was selected due to the evidence of immune responses, and particularly inflammation, being involved in cancer development, and due to the low cost of the array. This was the first large-scale genotyping platform to be used in South African OSCC samples. A case-control association study was performed using 278 cases and 257 controls (remaining from the original 300 of each after QC) from the Black population. Three SNPs were significantly associated with OSCC, using the Bonferroni correction to account for multiple testing ($P<1.84 \times 10^{-6}$). These variants were all located in *TGFBR3*, a co-receptor in the TGFβ signalling pathway which is involved in proliferation, apoptosis and immune responses (reviewed in Gatza *et al.* 2010). Due to insufficient sample numbers to complete an independent replication, an extension study was completed where an additional 126 cases and 577 controls were available. Seven SNPs were selected for genotyping in the extension study, which were genotyped in all samples available (407 OSCC cases and 834 controls). None of the variants were significantly associated with OSCC when accounting for multiple testing, with all but one variant becoming less significant in this analysis. In addition, no SNPs were associated with the disease when only the additional samples, those not genotyped on the Immunochip, were analyzed.

7.1.2 Somatic mutations in OSCC

Sequencing the exomes of 8 blood-tumour pairs identified between 10 and 301 somatic mutations in each tumour, although some of these are false positives. Somatic mutations were identified in several genes known to be mutated in cancer, including *TP53* and *KL*. However, 2 tumours did not contain any confirmed functional mutations. *TP53* and *PPM1D* were further investigated by Sanger sequencing the coding regions in all available blood-tumour pairs (10 and 11 pairs, respectively). *TP53* somatic mutations were identified in 60% of tumours. For *PPM1D*, only 1 tumour contained a somatic mutation (a stop codon), with four tumours showing evidence of loss of heterozygosity (LOH) at this region, suggesting that this may be an important alteration for OSCC development.

7.2 Lack of variants significantly associated with OSCC in South African populations

7.2.1 Candidate gene studies

In this thesis, a total of 18 variants previously associated with OSCC in Asian or European populations were tested for association with the disease in two South African populations. In the Mixed Ancestry population, two variants were significantly associated with the disease; *ALDH2* +82 A>G (rs886205) and *RUNX1* rs2014300. No variants were significantly, or suggestively ($P<0.05$), associated in the Black population.

A lack of replication for disease associated variants is fairly common, with several of the variants tested in this thesis previously showing inconsistent effects both within and between populations. This includes *CASP8* -652 6N indel (rs3834129) and *COX-2* -1195G>A (rs689466), which are both associated with OSCC in Chinese populations (Zhang *et al.* 2005; Sun *et al.* 2007) but not in a northern Indian population (Upadhyay *et al.* 2009; Umar *et al.* 2011).

Other studies have specifically focused on the effect of disease associated variants in different populations, and have found varying results. In a study of type 2 diabetes, 19 variants associated with the disease in a European population were tested for disease association in individuals from 5 other populations (European/American, African Americans, Latinos, Japanese Americans and Native Hawaiians) (Waters *et al.* 2010). The authors conclude that there were consistent associations for all risk variants across six populations, with the effect of the variants being in the same direction in the pooled analysis of all samples compared to the original European association. However, when observing the results for each population separately, the odds ratios were often in opposite directions to the pooled data, which suggests that some variants may have opposite effects, or have no effect, in different populations, which appears to contradict the authors' conclusions.

Ntzani *et al.* (2012) have compared allele frequencies and the effect sizes of 108 risk loci identified in association studies from different populations. This covered any disease where variants associated in GWAS had been tested for association in an additional population, with data for each variant obtained from two distinct ancestral groups; European, Asian or African. For African versus European populations, and African versus Asian studies, the genetic risk estimate was in the opposite direction or >2-fold different in 79% and 89% of studies, respectively. Given this large difference in effects, the authors conclude that it is not possible to predict the effect size of variants in different populations. However, the disparities, particularly when the genetic effect is in the same direction but with a >2-fold difference, may be due to underpowered studies in African populations. Indeed, Ntzani *et al.* note that studies in African populations showed weaker genetic effects than other populations, which is probably due to the GWAS data being obtained from the non-African populations. The reasons for this, in the context of the work in this thesis, are discussed below.

There are several explanations for the lack of replications observed in the South African populations. Firstly, the variants tested for association may have been false positives in the original European/Asian studies. However, with five of the variants originally identified in large GWAS which use independent replication phases, these are less likely to be false positives than associations identified through candidate gene association studies. In addition, most variants had been previously investigated in a number of populations and/or with different cancer subtypes.

Alternatively, risk loci may be population specific and may be absent in the South African populations. For example, the OSCC risk variant, *ALDH2* Glu504Lys (rs671), is specific to Asian populations (Li *et al.* 2009.a), and was monomorphic in the South African Black and Mixed Ancestry populations. However, the same gene may be associated with disease in several populations but be due to different risk variants. For example, *ALDH2* +82 A>G (rs886205) was associated with OSCC in the South African Mixed Ancestry population, showing a common role for *ALDH2* in OSCC susceptibility across populations. Therefore, even though a specific variant may not be associated with disease in a replication study, additional work may identify other variants in the same gene associated with disease across populations.

Perhaps the most important reason for the lack of replication observed in the South African study is due to the fact that association studies do not necessarily identify causal variants. The SNPs selected for genotyping in this thesis were all identified as OSCC susceptibility loci in European or Asian populations. These SNPs may be the causal variant or merely be in a high level of LD with it, i.e. tagging SNPs. Africans are known to be the most genetically diverse populations in the world, who have smaller haplotype blocks and lower levels of LD between variants (reviewed in Teo *et al.* 2010). Therefore, tagging SNPs associated with disease in European/Asian populations may be in very low LD with the causal variant in African populations. If this is the case, a disease

association would not be identified. If a moderate level of LD does exist between the causal and tagging variant in African populations, studies may lack power to detect the weaker effect size that would be observed. Indeed, it is known that it is more difficult to achieve genome-wide significance levels in African populations due to this problem (Jallow *et al.* 2009).

Lack of power may also be caused by inadequate sample sizes to detect a significance difference in allele frequency between cases and controls. This will be particularly relevant if the effect size of the risk variant is lower in the African populations compared to the original study, which may be due to genotyping of a tagging variant, as discussed above, or due to a genuine difference in effect size between populations.

Gene-gene or gene-environment interactions may also be important in OSCC susceptibility which may differ between populations. The latter of these was investigated in the South African populations for alcohol and tobacco use, but other population-specific environmental risk factors may exist.

7.2.2 Immunochip study

The lack of significant associations with OSCC in the Immunochip study could be the result of several factors. One is the small sample size, which was driven by the small budget for the project. A second is that the screen was restricted to loci with previous evidence of association with immune diseases and some additional phenotypes from the WTCCC. Thirdly, variants specific to African populations were not included as the Immunochip design was based on GWAS data from European populations. Therefore, it is possible regions that show suggestive evidence of association with OSCC harbour additional unknown variants which are more significantly associated with the disease. For example, with 4 out of the 5 most significantly associated variants located in *TGFBR3*, this gene may harbour further variants specific to African populations that were not investigated by the Immunochip. Sequencing of this gene in individuals from the

South African Black population and performing a case-control association analysis may identify novel variants which have a higher degree of association with OSCC. This approach has been used by Jallow *et al.* (2009), who fine-mapped a region associated with malaria resistance in their African GWAS. After imputing the data to a larger set of samples, they were able to identify a SNP more significantly associated with disease than any SNP on the GWAS panel in the same region. This variant was the known causal variant, providing evidence that this approach can be successful.

7.3 Advantages and disadvantages of genetic association studies in African populations

7.3.1 South African Black population

As discussed in section 7.2, African populations have the lowest levels of LD world-wide. This has proved problematic in genetic association studies, as unless the causal variants are genotyped directly, genetic associations might not be detected. However, using African populations can be a great advantage over other populations, since the lower level of LD can lead to the identification of causal variants (reviewed in Teo *et al.* 2010). To achieve this, the region of association needs to be fine-mapped to enable all polymorphic variants to be tested for association with the disease. It is also important to use a large set of SNPs to assess population structure in any population not previously studied in depth, and correct for this if necessary.

7.3.2 South African Mixed Ancestry population

The substantial amount of genetic heterogeneity in the composition of the South African Mixed Ancestry population leads to a series of questions when using these samples in genetic association studies. Should this population be used for candidate gene association studies? Would these samples be suitable for a GWAS? Is there a method to correct for the population structure?

In the candidate gene association studies carried out in this study, cases and controls were both obtained from the same region of the Western Cape in South Africa in order to limit heterogeneity between them. However, since only a small number of genetic variants were genotyped in these initial studies, there was insufficient data to assess population structure and apply appropriate corrections. Other studies have performed candidate gene case-control association studies without making any adjustments for population structure (de Wit *et al.* 2011; Matsha *et al.* 2012). De Wit *et al.* (2011) tested for gene-gene interactions between tuberculosis genes and did not adjust for population variation due to the population being relatively homogeneous. This conclusion was based on an earlier study that showed cases and controls from the Mixed Ancestry population not to be significantly different from each other, based on genotyping a panel of 25 SNPs (Barreiro *et al.* 2006). In contrast, another study found the Mixed Ancestry population to be substantially heterogeneous (Patterson *et al.* 2010), which is in agreement with the genotyping of our samples on the Immunochip. Patterson *et al.* (2010) recommend that all association studies in this population are corrected for population stratification, with EIGENSTRAT, which uses principal components analysis, being a suitable method (Price *et al.* 2006). However, this method requires a large number of SNPs to be genotyped. Therefore, the candidate gene association results in this thesis may contain both false-positive and false-negative results due to population stratification.

In future studies, large-scale genotyping platforms should be used for the South African Mixed Ancestry population, and it is proposed that genome-wide association studies can confidently be performed in this population using the EIGENSTRAT method for correction (Patterson *et al.* 2010). However, no studies have done so to date, although GWAS have been performed in other admixed populations, including African-Americans (Adeyemo *et al.* 2009; Lettre *et al.* 2011). Another method that has been successful in identifying disease susceptibility loci in admixed populations is by using admixture mapping

(reviewed in Winkler *et al.* 2010). This approach is based on the principle that both allele frequencies and disease incidence differ between populations. Individuals with a disease might have a chromosomal region that is more frequently inherited from the population ancestry that has a higher disease incidence. The strategy requires genotyping several thousand variants across the genome, which have different frequencies across populations. This could be achieved using ancestry-informative markers, a panel of SNPs known to differ in frequency amongst populations, or using genome-wide genotyping arrays. Most admixture mapping studies have been in the African-American populations, which are on average derived from 20% European ancestry and 80% African ancestry (Patterson *et al.* 2004; Reiner *et al.* 2005). No such studies have been performed in the South African Mixed Ancestry population, which may prove more complex to analyze due to the contribution of more ancestral groups.

7.4 Other genetic factors involved in disease susceptibility

This thesis investigated the association of common variants, mainly SNPs and 2 insertions/deletions, with susceptibility to OSCC. Other variants may also be involved in cancer susceptibility which were not investigated. Firstly, rare variants may contribute to susceptibility but sample numbers are required to be extremely large to identify significant associations, unless the effect sizes are very large. In addition, genotyping of a single SNP in a case-control association study is unlikely to be a successful approach. One recent method to investigate rare variants has been to utilise next generation sequencing for a panel of genes involved in DNA repair (Ruark *et al.* 2013). Focusing on protein-truncating variants (PTVs), *PPM1D* was shown to harbour more of these mutations in breast and ovarian cancer cases than controls. PTVs were present in 25 out of >7,000 cancer cases compared to 1 out of >5,000 controls, showing the large number of samples needed, which may be unattainable in many studies. The role of rare variants in OSCC has not been investigated in any population.

Copy number variants (CNVs) may also contribute to cancer susceptibility, as reviewed by Kuiper *et al.* (2010), although this has not been fully investigated due to more complex genotyping methods than SNP genotyping. However, it is unknown how much heritability CNVs will account for. A large study by the Wellcome Trust Case Control Consortium has shown that common CNVs make little contribution to common disease susceptibility (Craddock *et al.* 2010).

7.5 Somatic mutations

As discussed in detail in chapter 6, section 6.4, the results of the pilot whole-exome sequencing study in OSCC blood-tumour pairs show that substantial numbers of somatic mutations exist in most of these tumours, but their importance as potential driver mutations is, with the exception of *TP53*, not yet clear. This should be addressed by expanding exome sequencing to a larger number of blood-tumour pairs to look for recurrent mutation of genes, and following up the most promising findings by sequencing those genes in a large panel of tumours to determine the extent of their contribution to South African OSCC.

7.6 Limitations

The search for disease susceptibility genes in this thesis was largely restricted to the analysis of candidate-gene or GWAS-derived loci from studies in other populations, which is clearly a limitation. However, the negative results obtained are informative, since they suggest that there may be substantial differences in the genetic components of OSCC susceptibility particularly in the Black South African population. A further limitation is that this study does not rule out the possibility that other un-genotyped or as yet unknown variants in these genes are involved. However, a lack of genotyping platforms specifically designed for the South African populations does not permit a well-powered genome-wide approach for the identification of genetic susceptibility in these populations.

Another potential limitation for the genetic susceptibility studies was the available sample size. Sample sizes were increased during the course of the thesis, reaching 407 cases and 849 controls from the South African Black population, and 257 cases and 860 controls from the Mixed Ancestry population. Genome-wide association studies regularly use >2,000 cases and controls, so sample sizes will need to be increased to detect association of variants with moderate or small effect sizes.

The relatively low number of blood-tumour pairs that were exome sequenced in the pilot study (8 in total) limits the detection of genes that are recurrently mutated in OSCC, and more should be sequenced. Another limitation for this part of the project is that no potentially functional somatic mutations were detected in 2 OSCC tumours, suggesting that a high degree of normal tissue may have been present in the tumour samples. More preliminary quality control of tumour content is clearly required before submitting samples to exome sequencing.

7.7 Future directions

Directly following on from work described in this thesis, *TGFBR3* could be further investigated as a candidate gene, as the Immunochip data suggests that it might be involved in OSCC susceptibility. The first step would be to attempt to replicate the suggestive association by genotyping the top SNP in a further set of Black South African cases and controls, and then carry out a meta-analysis of the primary and replication data. If the analysis strengthened the evidence of association at this locus, ideally, the entire genomic region containing the gene (4550 bp) would be sequenced by targeted NGS in a panel of around 50-100 OSCC patients from the Black South African population to identify most common variants in this region and their LD relationships. This mini-Hapmap could then be used to select a subset of variants representing common variation across the gene for genotyping in a well-powered case-control association study.

For future OSCC association studies in the South African populations, we need to progress from the candidate gene approach. Given our knowledge of the lower levels of LD in African populations, and the inability to replicate disease associations throughout this thesis, it is evident that candidate gene studies testing one variant at a time, is not a suitable approach.

GWAS arrays specific to the South African Black and Mixed Ancestry populations are not currently available, and there is little or no whole-genome sequence data available. Therefore, it would be necessary to carry out whole-genome sequencing on a panel of individuals from these two populations and then design SNP arrays which provided good coverage of common variation. These could then be used to carry out well-powered GWAS for OSCC. An alternative strategy is to use low coverage whole-genome sequencing with imputation for the GWAS. This approach is based on the observation that even extremely low sequence coverage of 0.1-0.5x, when coupled with imputation to 1000 Genomes reference panels, can deliver high accuracy and power comparable to high density SNP arrays, and at a comparable or cheaper price (Pasaniuc *et al.* 2012). However, this is a new approach, and pilot data would be required to validate it.

Both of these methods are potentially powerful to identify variants associated with OSCC in the South African study, as they are genome-wide and would address the issue of the lower LD levels in African populations. Apart from the Immunochip study, comprehensive studies have not been performed in the South African population, and therefore, the variants with the largest genetic effects (if present), the ‘low-hanging fruit’, could potentially be detected with a moderate sample size. However, as other association studies have shown, sample numbers are required to be large to have enough power to identify common variants of low-moderate effect. Therefore, sample sizes must continue

to be increased in order to perform well-powered studies. Most GWAS now genotype at least 2,000 cases and controls.

As discussed in section 7.5, the whole exome pilot sequencing project to detect somatic mutations in OSCC could be followed up at relatively modest cost by exome sequencing a larger number of blood-tumour pairs and following up recurrently mutated genes in a larger panel of tumours. However, it is clear that many important sequence changes such as large-scale rearrangements or copy number changes and mutations in regulatory regions would be missed using the exome approach. In future, therefore, a substantial number of blood-tumour pairs should be whole-genome sequenced to identify all possible somatic variants, as recommended by the International Cancer Genome Consortium (www.icgc.org).

7.8 Conclusions

The aim of this thesis was to investigate genetic susceptibility to oesophageal squamous cell carcinoma in the Black and Mixed Ancestry populations of South Africa. Based on candidate-gene association studies, one SNP, *PLCE1* Arg548Leu (rs17417407), was significantly associated with OSCC in the Black population, and two variants, *ALDH2* +82 A>G (rs886205) and *RUNX1* rs2014300 were associated in the Mixed Ancestry population. Another 15-20 variants were found to be not associated with the disease in each population, which suggests that there may be substantial differences in the genetic architecture of OSCC in African populations. The finding in our Immunochip study that multiple variants in the *TGFBR3* gene show suggestive association with OSCC is promising, as the TGFβ pathway has been shown to be important in preventing OSCC (Achyut *et al.* 2013).

These studies highlight the need for genotyping arrays that are designed for the populations that are to be tested, or the development of targeted next generation sequencing approaches. High throughput genomic infrastructure is

not available in most African countries at present, but collaborations and capacity building via, for example, the H3 Africa Genetics programme, could address this (h3africa.org).

Finally, the pilot exome sequencing project to screen for somatic mutations in tumours from OSCC patients from South African populations identified mutations present in genes already known to be mutated in other cancers, including tumour suppressor genes. *TP53* was the most commonly mutated gene, with mutations occurring in 60% of tumours. This study should be expanded in order to identify likely driver mutations in OSCC. This is important to understand the mechanisms of OSCC development, which may lead to the identification of drug targets and identify patients that are likely to respond to specific treatments.

References

Abnet, C. C., Freedman, N. D., Hu, N., Wang, Z., Yu, K., Shu, X. O., Yuan, J. M., Zheng, W., Dawsey, S. M., Dong, L. M., Lee, M. P., Ding, T., Qiao, Y. L., Gao, Y. T., Koh, W. P., Xiang, Y. B., Tang, Z. Z., Fan, J. H., Wang, C., Wheeler, W., Gail, M. H., Yeager, M., Yuenger, J., Hutchinson, A., Jacobs, K. B., Giffen, C. A., Burdett, L., Fraumeni, J. F., Jr., Tucker, M. A., Chow, W. H., Goldstein, A. M., Chanock, S. J. and Taylor, P. R. (2010). "A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma." Nat Genet **42**(9): 764-767.

Abnet, C. C., Wang, Z., Song, X., Hu, N., Zhou, F. Y., Freedman, N. D., Li, X. M., Yu, K., Shu, X. O., Yuan, J. M., Zheng, W., Dawsey, S. M., Liao, L. M., Lee, M. P., Ding, T., Qiao, Y. L., Gao, Y. T., Koh, W. P., Xiang, Y. B., Tang, Z. Z., Fan, J. H., Chung, C. C., Wang, C., Wheeler, W., Yeager, M., Yuenger, J., Hutchinson, A., Jacobs, K. B., Giffen, C. A., Burdett, L., Fraumeni, J. F., Jr., Tucker, M. A., Chow, W. H., Zhao, X. K., Li, J. M., Li, A. L., Sun, L. D., Wei, W., Li, J. L., Zhang, P., Li, H. L., Cui, W. Y., Wang, W. P., Liu, Z. C., Yang, X., Fu, W. J., Cui, J. L., Lin, H. L., Zhu, W. L., Liu, M., Chen, X., Chen, J., Guo, L., Han, J. J., Zhou, S. L., Huang, J., Wu, Y., Yuan, C., Ji, A. F., Kul, J. W., Fan, Z. M., Wang, J. P., Zhang, D. Y., Zhang, L. Q., Zhang, W., Chen, Y. F., Ren, J. L., Dong, J. C., Xing, G. L., Guo, Z. G., Yang, J. X., Mao, Y. M., Yuan, Y., Guo, E. T., Hou, Z. C., Liu, J., Li, Y., Tang, S., Chang, J., Peng, X. Q., Han, M., Yin, W. L., Liu, Y. L., Hu, Y. L., Liu, Y., Yang, L. Q., Zhu, F. G., Yang, X. F., Feng, X. S., Gao, S. G., Liu, H. L., Yuan, L., Jin, Y., Zhang, Y. R., Sheyhidin, I., Li, F., Chen, B. P., Ren, S. W., Liu, B., Li, D., Zhang, G. F., Yue, W. B., Feng, C. W., Qige, Q., Zhao, J. T., Yang, W. J., Lei, G. Y., Chen, L. Q., Li, E. M., Xu, L. Y., Wu, Z. Y., Bao, Z. Q., Chen, J. L., Li, X. C., Zhuang, X., Zhou, Y. F., Zuo, X. B., Dong, Z. M., Wang, L. W., Fan, X. P., Wang, J., Zhou, Q., Ma, G. S., Zhang, Q. X., Liu, H., Jian, X. Y., Lian, S. Y., Wang, J. S., Chang, F. B., Lu, C. D., Miao, J. J., Chen, Z. G., Wang, R., Guo, M., Fan, Z. L., Tao, P., Liu, T. J., Wei, J. C., Kong, Q. P., Fan, L., Wang, X. Z., Gao, F. S., Wang, T. Y., Xie, D., Wang, L., Chen, S. Q., Yang, W. C., Hong, J. Y., Qiu, S. L., Goldstein, A. M., Yuan, Z. Q., Chanock, S. J., Zhang, X. J., Taylor, P. R. and Wang, L. D. (2012). "Genotypic variants at 2q33 and risk of esophageal squamous cell carcinoma in China: a meta-analysis of genome-wide association studies." Hum Mol Genet **21**(9): 2132-2141.

Achyut, B. R., Bader, D. A., Robles, A. I., Wangsa, D., Harris, C. C., Ried, T. and Yang, L. (2013). "Inflammation-mediated genetic and epigenetic alterations drive cancer development in the neighboring epithelium upon stromal abrogation of TGF-beta signaling." PLoS Genet **9**(2): e1003251.

Adams, J. M. and Cory, S. (1998). "The Bcl-2 protein family: arbiters of cell survival." Science **281**(5381): 1322-1326.

Adeyemo, A., Gerry, N., Chen, G., Herbert, A., Doumatey, A., Huang, H., Zhou, J., Lashley, K., Chen, Y., Christman, M. and Rotimi, C. (2009). "A genome-wide association study of hypertension and blood pressure in African Americans." PLoS Genet **5**(7): e1000564.

Agrawal, N., Frederick, M. J., Pickering, C. R., Bettegowda, C., Chang, K., Li, R. J., Fakhry, C., Xie, T. X., Zhang, J., Wang, J., Zhang, N., El-Naggar, A. K., Jasser, S. A., Weinstein, J. N., Trevino, L., Drummond, J. A., Muzny, D. M., Wu, Y., Wood, L. D., Hruban, R. H., Westra, W. H., Koch, W. M., Califano, J. A., Gibbs, R. A., Sidransky, D., Vogelstein, B., Velculescu, V. E., Papadopoulos, N., Wheeler, D. A., Kinzler, K. W. and Myers, J. N. (2011). "Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1." Science **333**(6046): 1154-1157.

Agrawal, N., Jiao, Y., Bettegowda, C., Hutfless, S. M., Wang, Y., David, S., Cheng, Y., Twaddell, W. S., Latt, N. L., Shin, E. J., Wang, L. D., Wang, L., Yang, W., Velculescu, V. E., Vogelstein, B., Papadopoulos, N., Kinzler, K. W. and Meltzer, S. J. (2012). "Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma." Cancer Discov **2**(10): 899-905.

Akbari, M. R., Malekzadeh, R., Nasrollahzadeh, D., Amanian, D., Sun, P., Islami, F., Sotoudeh, M., Semnani, S., Boffeta, P., Dawsey, S. M., Ghadirian, P. and Narod, S. A. (2006). "Familial risks of esophageal cancer among the Turkmen population of the Caspian littoral of Iran." Int J Cancer **119**(5): 1047-1051.

Akbari, M. R., Malekzadeh, R., Shakeri, R., Nasrollahzadeh, D., Foumani, M., Sun, Y., Pourshams, A., Sadjadi, A., Jafari, E., Sotoudeh, M., Kamangar, F., Boffetta, P., Dawsey, S. M., Ghadirian, P. and Narod, S. A. (2009). "Candidate gene association study of esophageal squamous cell carcinoma in a high-risk region in Iran." Cancer Res **69**(20): 7994-8000.

Anderson, L. A., Johnston, B. T., Watson, R. G., Murphy, S. J., Ferguson, H. R., Comber, H., McGuigan, J., Reynolds, J. V. and Murray, L. J. (2006). "Nonsteroidal anti-inflammatory drugs and the esophageal inflammation-metaplasia-adenocarcinoma sequence." Cancer Res **66**(9): 4975-4982.

Balkwill, F. and Mantovani, A. (2001). "Inflammation and cancer: back to Virchow?" Lancet **357**(9255): 539-545.

Barreiro, L. B., Neyrolles, O., Babb, C. L., Tailleux, L., Quach, H., McElreavey, K., Helden, P. D., Hoal, E. G., Gicquel, B. and Quintana-Murci, L. (2006). "Promoter variation in the DC-SIGN-encoding gene CD209 is associated with tuberculosis." PLoS Med **3**(2): e20.

Barrett, J. C., Fry, B., Maller, J. and Daly, M. J. (2005). "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics **21**(2): 263-265.

Bell, D. W., Sikdar, N., Lee, K. Y., Price, J. C., Chatterjee, R., Park, H. D., Fox, J., Ishiai, M., Rudd, M. L., Pollock, L. M., Fogoros, S. K., Mohamed, H., Hanigan, C. L., Zhang, S. Y., Cruz, P., Renaud, G., Hansen, N. F., Cherukuri, P. F., Borate, B., McManus, K. J., Stoepel, J., Sipahimalani, P., Godwin, A. K., Sgroi, D. C., Merino, M. J., Elliot, G., Elkahloun, A., Vinson, C., Takata, M., Mullikin, J. C., Wolfsberg, T. G., Hieter, P., Lim, D. S. and Myung, K. (2011). "Predisposition to cancer caused by genetic and functional defects of mammalian Atad5." PLoS Genet **7**(8): e1002245.

Benn, M., Tybjaerg-Hansen, A., Stender, S., Frikke-Schmidt, R. and Nordestgaard, B. G. (2011). "Low-density lipoprotein cholesterol and the risk of cancer: a mendelian randomization study." J Natl Cancer Inst **103**(6): 508-519.

Berger, A. H., Knudson, A. G. and Pandolfi, P. P. (2011). "A continuum model for tumour suppression." Nature **476**(7359): 163-169.

Bernabeu, C., Lopez-Novoa, J. M. and Quintanilla, M. (2009). "The emerging role of TGF-beta superfamily coreceptors in cancer." Biochim Biophys Acta **1792**(10): 954-973.

Biankin, A. V., Waddell, N., Kassahn, K. S., Gingras, M. C., Muthuswamy, L. B., Johns, A. L., Miller, D. K., Wilson, P. J., Patch, A. M., Wu, J., Chang, D. K., Cowley, M. J., Gardiner, B. B., Song, S., Harliwong, I., Idrisoglu, S., Nourse, C., Nourbakhsh, E., Manning, S., Wani, S., Gongora, M., Pajic, M., Scarlett, C. J., Gill, A. J., Pinho, A. V., Rooman, I., Anderson, M., Holmes, O., Leonard, C., Taylor, D., Wood, S., Xu, Q., Nones, K., Fink, J. L., Christ, A., Bruxner, T., Cloonan, N., Kolle, G., Newell, F., Pinese, M., Mead, R. S., Humphris, J. L., Kaplan, W., Jones, M. D., Colvin, E. K., Nagrial, A. M., Humphrey, E. S., Chou, A., Chin, V. T., Chantrill, L. A., Mawson, A., Samra, J. S., Kench, J. G., Lovell, J. A., Daly, R. J., Merrett, N. D., Toon, C., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Kakkar, N., Zhao, F., Wu, Y. Q., Wang, M., Muzny, D. M., Fisher, W. E., Brunicardi, F. C., Hodges, S. E., Reid, J. G., Drummond, J., Chang, K., Han, Y., Lewis, L. R., Dinh, H., Buhay, C. J., Beck, T., Timms, L., Sam, M., Begley, K., Brown, A., Pai, D., Panchal, A., Buchner, N., De Borja, R., Denroche, R. E., Yung, C. K., Serra, S., Onetto, N., Mukhopadhyay, D.,

Tsao, M. S., Shaw, P. A., Petersen, G. M., Gallinger, S., Hruban, R. H., Maitra, A., Iacobuzio-Donahue, C. A., Schulick, R. D., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Capelli, P., Corbo, V., Scardoni, M., Tortora, G., Tempero, M. A., Mann, K. M., Jenkins, N. A., Perez-Mancera, P. A., Adams, D. J., Largaespada, D. A., Wessels, L. F., Rust, A. G., Stein, L. D., Tuveson, D. A., Copeland, N. G., Musgrove, E. A., Scarpa, A., Eshleman, J. R., Hudson, T. J., Sutherland, R. L., Wheeler, D. A., Pearson, J. V., McPherson, J. D., Gibbs, R. A. and Grimmond, S. M. (2012). "Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes." Nature **491**(7424): 399-405.

Blot, W. J., Li, J. Y., Taylor, P. R., Guo, W., Dawsey, S., Wang, G. Q., Yang, C. S., Zheng, S. F., Gail, M., Li, G. Y., Yu, Y., Liu, B. Q., Tangrea, J., Sun, Y. H., Liu, F., Fraumeni, J. F., Jr., Zhang, Y. H. and Li, B. (1993). "Nutrition intervention trials in Linxian, China: supplementation with specific vitamin/mineral combinations, cancer incidence, and disease-specific mortality in the general population." J Natl Cancer Inst **85**(18): 1483-1492.

Blot, W. J. and McLaughlin, J. K. (1999). "The changing epidemiology of esophageal cancer." Semin Oncol **26**(5 Suppl 15): 2-8.

Blyth, K., Cameron, E. R. and Neil, J. C. (2005). "The RUNX genes: gain or loss of function in cancer." Nat Rev Cancer **5**(5): 376-387.

Boffetta, P., Couto, E., Wichmann, J., Ferrari, P., Trichopoulos, D., Bueno-de-Mesquita, H. B., van Duijnhoven, F. J., Buchner, F. L., Key, T., Boeing, H., Nothlings, U., Linseisen, J., Gonzalez, C. A., Overvad, K., Nielsen, M. R., Tjonneland, A., Olsen, A., Clavel-Chapelon, F., Boutron-Ruault, M. C., Morois, S., Lagiou, P., Naska, A., Benetou, V., Kaaks, R., Rohrmann, S., Panico, S., Sieri, S., Vineis, P., Palli, D., van Gils, C. H., Peeters, P. H., Lund, E., Brustad, M., Engeset, D., Huerta, J. M., Rodriguez, L., Sanchez, M. J., Dorronsoro, M., Barricarte, A., Hallmans, G., Johansson, I., Manjer, J., Sonestedt, E., Allen, N. E., Bingham, S., Khaw, K. T., Slimani, N., Jenab, M., Mouw, T., Norat, T., Riboli, E. and Trichopoulou, A. (2010). "Fruit and vegetable intake and overall cancer risk in the European Prospective Investigation Into Cancer and Nutrition (EPIC)." J Natl Cancer Inst **102**(8): 529-537.

Bollschweiler, E., Wolfgarten, E., Gutschow, C. and Holscher, A. H. (2001). "Demographic variations in the rising incidence of esophageal adenocarcinoma in white males." Cancer **92**(3): 549-555.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M. and

Snyder, M. (2012). "Annotation of functional variation in personal genomes using RegulomeDB." Genome Res **22**(9): 1790-1797.

Brooks, P. J., Enoch, M. A., Goldman, D., Li, T. K. and Yokoyama, A. (2009.a). "The alcohol flushing response: an unrecognized risk factor for esophageal cancer from alcohol consumption." PLoS Med **6**(3): e50.

Brooks, P. J., Goldman, D. and Li, T. K. (2009.b). "Alleles of alcohol and acetaldehyde metabolism genes modulate susceptibility to oesophageal cancer from alcohol consumption." Hum Genomics **3**(2): 103-105.

Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., Froment, A., Bodo, J. M., Wambebe, C., Tishkoff, S. A. and Bustamante, C. D. (2010). "Genome-wide patterns of population structure and admixture in West Africans and African Americans." Proc Natl Acad Sci U S A **107**(2): 786-791.

Bulavin, D. V., Demidov, O. N., Saito, S., Kauraniemi, P., Phillips, C., Amundson, S. A., Ambrosino, C., Sauter, G., Nebreda, A. R., Anderson, C. W., Kallioniemi, A., Fornace, A. J., Jr. and Appella, E. (2002). "Amplification of PPM1D in human tumors abrogates p53 tumor-suppressor activity." Nat Genet **31**(2): 210-215.

Bulavin, D. V., Phillips, C., Nannenga, B., Timofeev, O., Donehower, L. A., Anderson, C. W., Appella, E. and Fornace, A. J., Jr. (2004). "Inactivation of the Wip1 phosphatase inhibits mammary tumorigenesis through p38 MAPK-mediated activation of the p16(Ink4a)-p19(Arf) pathway." Nat Genet **36**(4): 343-350.

Burrell, R. J. (1957). "Oesophageal cancer in the Bantu." S Afr Med J **31**(17): 401-409.

Burrell, R. J. (1962). "Esophageal cancer among Bantu in the Transkei." J Natl Cancer Inst **28**: 495-514.

Burrell, R. J. (1969). "Distribution maps of esophageal cancer among Bantu in the Transkei." J Natl Cancer Inst **43**(4): 877-889.

Burrell, R. J., Roach, W. A. and Shadwell, A. (1966). "Esophageal cancer in the Bantu of the Transkei associated with mineral deficiency in garden plants." J Natl Cancer Inst **36**(2): 201-209.

Bye, H., Prescott, N. J., Lewis, C. M., Matejcic, M., Moodley, L., Robertson, B., Rensburg, C., Parker, M. I. and Mathew, C. G. (2012). "Distinct genetic association at the PLCE1 locus with oesophageal squamous cell

carcinoma in the South African population." Carcinogenesis **33**(11): 2155-2161.

Bye, H., Prescott, N. J., Matejcic, M., Rose, E., Lewis, C. M., Parker, M. I. and Mathew, C. G. (2011). "Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa." Carcinogenesis **32**(12): 1855-1861.

Campbell, M. C. and Tishkoff, S. A. (2008). "African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping." Annu Rev Genomics Hum Genet **9**: 403-433.

Campbell, M. C. and Tishkoff, S. A. (2010). "The evolution of human genetic and phenotypic variation in Africa." Curr Biol **20**(4): R166-173.

Cao, Y. and Prescott, S. M. (2002). "Many actions of cyclooxygenase-2 in cellular dynamics and in cancer." J Cell Physiol **190**(3): 279-286.

Castellsague, X., Munoz, N., De Stefani, E., Victora, C. G., Castelletto, R. and Rolon, P. A. (2000). "Influence of mate drinking, hot beverages and diet on esophageal cancer risk in South America." Int J Cancer **88**(4): 658-664.

Ceol, C. J., Houvras, Y., Jane-Valbuena, J., Bilodeau, S., Orlando, D. A., Battisti, V., Fritsch, L., Lin, W. M., Hollmann, T. J., Ferre, F., Bourque, C., Burke, C. J., Turner, L., Uong, A., Johnson, L. A., Beroukhim, R., Mermel, C. H., Loda, M., Ait-Si-Ali, S., Garraway, L. A., Young, R. A. and Zon, L. I. (2011). "The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset." Nature **471**(7339): 513-517.

Chang-Claude, J., Becher, H., Blettner, M., Qiu, S., Yang, G. and Wahrendorf, J. (1997). "Familial aggregation of oesophageal cancer in a high incidence area in China." Int J Epidemiol **26**(6): 1159-1165.

Chattopadhyay, I., Singh, A., Phukan, R., Purkayastha, J., Kataki, A., Mahanta, J., Saxena, S. and Kapur, S. (2010). "Genome-wide analysis of chromosomal alterations in patients with esophageal squamous cell carcinoma exposed to tobacco and betel quid from high-risk area in India." Mutat Res **696**(2): 130-138.

Chen, L., Chen, D. T., Kurtyka, C., Rawal, B., Fulp, W. J., Haura, E. B. and Cress, W. D. (2012.b). "Tripartite motif containing 28 (Trim28) can regulate cell proliferation by bridging HDAC1/E2F interactions." J Biol Chem **287**(48): 40106-40118.

Chen, L., Park, S. M., Tumanov, A. V., Hau, A., Sawada, K., Feig, C., Turner, J. R., Fu, Y. X., Romero, I. L., Lengyel, E. and Peter, M. E. (2010). "CD95 promotes tumour growth." Nature **465**(7297): 492-496.

Chen, Y., Gao, X. J., Deng, Y. C. and Zhang, H. X. (2012.a). "Relationship between HLA-G gene polymorphism and the susceptibility of esophageal cancer in Kazakh and Han nationality in Xinjiang." Biomarkers **17**(1): 9-15.

Cheung, W. Y. and Liu, G. (2009). "Genetic variations in esophageal cancer risk and prognosis." Gastroenterol Clin North Am **38**(1): 75-91.

Chu, H., Cao, W., Chen, W., Pan, S., Xiao, Y., Liu, Y., Gu, H., Guo, W., Xu, L., Hu, Z. and Shen, H. (2012). "Potentially functional polymorphisms in IL-23 receptor and risk of esophageal cancer in a Chinese population." Int J Cancer **130**(5): 1093-1097.

Cook, M. B., Chow, W. H. and Devesa, S. S. (2009). "Oesophageal cancer incidence in the United States by race, sex, and histologic type, 1977-2005." Br J Cancer **101**(5): 855-859.

Cortes, A. and Brown, M. A. (2011). "Promise and pitfalls of the Immunochip." Arthritis Res Ther **13**(1): 101.

Cox, A., Dunning, A. M., Garcia-Closas, M., Balasubramanian, S., Reed, M. W., Pooley, K. A., Scollen, S., Baynes, C., Ponder, B. A., Chanock, S., Lissowska, J., Brinton, L., Peplonska, B., Southey, M. C., Hopper, J. L., McCredie, M. R., Giles, G. G., Fletcher, O., Johnson, N., dos Santos Silva, I., Gibson, L., Bojesen, S. E., Nordestgaard, B. G., Axelsson, C. K., Torres, D., Hamann, U., Justenhoven, C., Brauch, H., Chang-Claude, J., Kropp, S., Risch, A., Wang-Gohrke, S., Schurmann, P., Bogdanova, N., Dork, T., Fagerholm, R., Aaltonen, K., Blomqvist, C., Nevanlinna, H., Seal, S., Renwick, A., Stratton, M. R., Rahman, N., Sangrajrang, S., Hughes, D., Odefrey, F., Brennan, P., Spurdle, A. B., Chenevix-Trench, G., Beesley, J., Mannermaa, A., Hartikainen, J., Kataja, V., Kosma, V. M., Couch, F. J., Olson, J. E., Goode, E. L., Broeks, A., Schmidt, M. K., Hogervorst, F. B., Van't Veer, L. J., Kang, D., Yoo, K. Y., Noh, D. Y., Ahn, S. H., Wedren, S., Hall, P., Low, Y. L., Liu, J., Milne, R. L., Ribas, G., Gonzalez-Neira, A., Benitez, J., Sigurdson, A. J., Stredrick, D. L., Alexander, B. H., Struewing, J. P., Pharoah, P. D. and Easton, D. F. (2007). "A common coding variant in CASP8 is associated with breast cancer risk." Nat Genet **39**(3): 352-358.

Crabb, D. W., Edenberg, H. J., Bosron, W. F. and Li, T. K. (1989). "Genotypes for aldehyde dehydrogenase deficiency and alcohol sensitivity. The inactive ALDH2(2) allele is dominant." J Clin Invest **83**(1): 314-316.

Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D. F., Giannoulatou, E., Holmes, C., Marchini, J. L., Stirrups, K., Tobin, M. D., Wain, L. V., Yau, C., Aerts, J., Ahmad, T., Andrews, T. D., Arbury, H., Attwood, A., Auton, A., Ball, S. G., Balmforth, A. J., Barrett, J. C., Barroso, I., Barton, A., Bennett, A. J., Bhaskar, S., Blaszczyk, K., Bowes, J., Brand, O. J., Braund, P. S., Bredin, F., Breen, G., Brown, M. J., Bruce, I. N., Bull, J., Burren, O. S., Burton, J., Byrnes, J., Caesar, S., Clee, C. M., Coffey, A. J., Connell, J. M., Cooper, J. D., Dominiczak, A. F., Downes, K., Drummond, H. E., Dudakia, D., Dunham, A., Ebbs, B., Eccles, D., Edkins, S., Edwards, C., Elliot, A., Emery, P., Evans, D. M., Evans, G., Eyre, S., Farmer, A., Ferrier, I. N., Feuk, L., Fitzgerald, T., Flynn, E., Forbes, A., Forty, L., Franklyn, J. A., Freathy, R. M., Gibbs, P., Gilbert, P., Gokumen, O., Gordon-Smith, K., Gray, E., Green, E., Groves, C. J., Grozeva, D., Gwilliam, R., Hall, A., Hammond, N., Hardy, M., Harrison, P., Hassanali, N., Hebaishi, H., Hines, S., Hinks, A., Hitman, G. A., Hocking, L., Howard, E., Howard, P., Howson, J. M., Hughes, D., Hunt, S., Isaacs, J. D., Jain, M., Jewell, D. P., Johnson, T., Jolley, J. D., Jones, I. R., Jones, L. A., Kirov, G., Langford, C. F., Lango-Allen, H., Lathrop, G. M., Lee, J., Lee, K. L., Lees, C., Lewis, K., Lindgren, C. M., Maisuria-Armer, M., Maller, J., Mansfield, J., Martin, P., Massey, D. C., McArdle, W. L., McGuffin, P., McLay, K. E., Mentzer, A., Mimmack, M. L., Morgan, A. E., Morris, A. P., Mowat, C., Myers, S., Newman, W., Nimmo, E. R., O'Donovan, M. C., Onipinla, A., Onyiah, I., Ovington, N. R., Owen, M. J., Palin, K., Parnell, K., Pernet, D., Perry, J. R., Phillips, A., Pinto, D., Prescott, N. J., Prokopenko, I., Quail, M. A., Rafelt, S., Rayner, N. W., Redon, R., Reid, D. M., Renwick, Ring, S. M., Robertson, N., Russell, E., St Clair, D., Sambrook, J. G., Sanderson, J. D., Schuilenburg, H., Scott, C. E., Scott, R., Seal, S., Shaw-Hawkins, S., Shields, B. M., Simmonds, M. J., Smyth, D. J., Somaskantharajah, E., Spanova, K., Steer, S., Stephens, J., Stevens, H. E., Stone, M. A., Su, Z., Symmons, D. P., Thompson, J. R., Thomson, W., Travers, M. E., Turnbull, C., Valsesia, A., Walker, M., Walker, N. M., Wallace, C., Warren-Perry, M., Watkins, N. A., Webster, J., Weedon, M. N., Wilson, A. G., Woodburn, M., Wordsworth, B. P., Young, A. H., Zeggini, E., Carter, N. P., Frayling, T. M., Lee, C., McVean, G., Munroe, P. B., Palotie, A., Sawcer, S. J., Scherer, S. W., Strachan, D. P., Tyler-Smith, C., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Gough, S. C., Hall, A. S., Hattersley, A. T., Hill, A. V., Mathew, C. G., Pembrey, M., Satsangi, J., Stratton, M. R., Worthington, J., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W., Parkes, M., Rahman, N., Todd, J. A., Samani, N. J. and Donnelly, P. (2010). "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls." Nature **464**(7289): 713-720.

Cui, R., Kamatani, Y., Takahashi, A., Usami, M., Hosono, N., Kawaguchi, T., Tsunoda, T., Kamatani, N., Kubo, M., Nakamura, Y. and Matsuda, K. (2009). "Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk." Gastroenterology **137**(5): 1768-1775.

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Graf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerod, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Borresen-Dale, A. L., Brenton, J. D., Tavare, S., Caldas, C. and Aparicio, S. (2012). "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups." Nature **486**(7403): 346-352.

Dandara, C., Ballo, R. and Parker, M. I. (2005). "CYP3A5 genotypes and risk of oesophageal cancer in two South African populations." Cancer Lett **225**(2): 275-282.

Dandara, C., Li, D. P., Walther, G. and Parker, M. I. (2006). "Gene-environment interaction: the role of SULT1A1 and CYP3A5 polymorphisms as risk modifiers for squamous cell carcinoma of the oesophagus." Carcinogenesis **27**(4): 791-797.

Danielsen, S. A., Cekaite, L., Agesen, T. H., Sveen, A., Nesbakken, A., Thiis-Evensen, E., Skotheim, R. I., Lind, G. E. and Lothe, R. A. (2011). "Phospholipase C isozymes are deregulated in colorectal cancer--insights gained from gene set enrichment analysis of the transcriptome." PLoS One **6**(9): e24419.

De Bont, R. and van Larebeke, N. (2004). "Endogenous DNA damage in humans: a review of quantitative data." Mutagenesis **19**(3): 169-185.

de Martel, C., Ferlay, J., Franceschi, S., Vignat, J., Bray, F., Forman, D. and Plummer, M. (2012). "Global burden of cancers attributable to infections in 2008: a review and synthetic analysis." Lancet Oncol **13**(6): 607-615.

De Stefani, E., Deneo-Pellegrini, H., Ronco, A. L., Boffetta, P., Brennan, P., Munoz, N., Castellsague, X., Correa, P. and Mendilaharsu, M. (2003). "Food groups and risk of squamous cell carcinoma of the oesophagus: a case-control study in Uruguay." Br J Cancer **89**(7): 1209-1214.

de Visser, K. E., Eichten, A. and Coussens, L. M. (2006). "Paradoxical roles of the immune system during cancer development." Nat Rev Cancer **6**(1): 24-37.

de Wit, E., Delport, W., Rugamika, C. E., Meintjes, A., Moller, M., van Helden, P. D., Seoighe, C. and Hoal, E. G. (2010). "Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape." Hum Genet **128**(2): 145-153.

de Wit, E., van der Merwe, L., van Helden, P. D. and Hoal, E. G. (2011). "Gene-gene interaction between tuberculosis candidate genes in a South African population." Mamm Genome **22**(1-2): 100-110.

Deere, H. M. R. (2007). Pathology of Esophageal Cancer. Carcinoma of the Esophagus. Rankin, S. C. Cambridge, UK, Cambridge University Press

Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T. L., Thompson, J. R., Ingelsson, E., Saleheen, D., Erdmann, J., Goldstein, B. A., Stirrups, K., Konig, I. R., Cazier, J. B., Johansson, A., Hall, A. S., Lee, J. Y., Willer, C. J., Chambers, J. C., Esko, T., Folkersen, L., Goel, A., Grundberg, E., Havulinna, A. S., Ho, W. K., Hopewell, J. C., Eriksson, N., Kleber, M. E., Kristiansson, K., Lundmark, P., Lyytikainen, L. P., Rafelt, S., Shungin, D., Strawbridge, R. J., Thorleifsson, G., Tikkanen, E., Van Zuydam, N., Voight, B. F., Waite, L. L., Zhang, W., Ziegler, A., Absher, D., Altshuler, D., Balmforth, A. J., Barroso, I., Braund, P. S., Burgdorf, C., Claudi-Boehm, S., Cox, D., Dimitriou, M., Do, R., Doney, A. S., Mokhtari, N. E., Eriksson, P., Fischer, K., Fontanillas, P., Franco-Cereceda, A., Gigante, B., Groop, L., Gustafsson, S., Hager, J., Hallmans, G., Han, B. G., Hunt, S. E., Kang, H. M., Illig, T., Kessler, T., Knowles, J. W., Kolovou, G., Kuusisto, J., Langenberg, C., Langford, C., Leander, K., Lokki, M. L., Lundmark, A., McCarthy, M. I., Meisinger, C., Melander, O., Mihailov, E., Maouche, S., Morris, A. D., Muller-Nurasyid, M., Nikus, K., Peden, J. F., Rayner, N. W., Rasheed, A., Rosinger, S., Rubin, D., Rumpf, M. P., Schafer, A., Sivananthan, M., Song, C., Stewart, A. F., Tan, S. T., Thorgeirsson, G., Schoot, C. E., Wagner, P. J., Wells, G. A., Wild, P. S., Yang, T. P., Amouyel, P., Arveiler, D., Basart, H., Boehnke, M., Boerwinkle, E., Brambilla, P., Cambien, F., Cupples, A. L., de Faire, U., Dehghan, A., Diemert, P., Epstein, S. E., Evans, A., Ferrario, M. M., Ferrieres, J., Gauguier, D., Go, A. S., Goodall, A. H., Gudnason, V., Hazen, S. L., Holm, H., Iribarren, C., Jang, Y., Kahonen, M., Kee, F., Kim, H. S., Klopp, N., Koenig, W., Kratzer, W., Kuulasmaa, K., Laakso, M., Laaksonen, R., Lind, L., Ouwehand, W. H., Parish, S., Park, J. E., Pedersen, N. L., Peters, A., Quertermous, T., Rader, D. J., Salomaa, V., Schadt, E., Shah, S. H., Sinisalo, J., Stark, K., Stefansson, K., Tregouet, D. A., Virtamo, J., Wallentin, L., Wareham, N., Zimmermann, M. E., Nieminen, M. S., Hengstenberg, C., Sandhu, M. S., Pastinen, T., Syvanen, A. C., Hovingh, G. K., Dedoussis, G., Franks, P. W., Lehtimaki, T., Metspalu, A., Zalloua, P. A., Siegbahn, A., Schreiber, S., Ripatti, S., Blankenberg, S. S., Perola, M., Clarke, R., Boehm, B. O., O'Donnell, C., Reilly, M. P., Marz, W., Collins, R., Kathiresan, S., Hamsten, A., Kooner,

J. S., Thorsteinsdottir, U., Danesh, J., Palmer, C. N., Roberts, R., Watkins, H., Schunkert, H. and Samani, N. J. (2012). "Large-scale association analysis identifies new risk loci for coronary artery disease." Nat Genet **45**(1): 25-33.

Ding, J. H., Li, S. P., Cao, H. X., Wu, J. Z., Gao, C. M., Liu, Y. T., Zhou, J. N., Chang, J. and Yao, G. H. (2009). "Alcohol dehydrogenase-2 and aldehyde dehydrogenase-2 genotypes, alcohol drinking and the risk for esophageal cancer in a Chinese population." J Hum Genet **55**(2): 97-102.

Dobreva, G., Chahrour, M., Dautzenberg, M., Chirivella, L., Kanzler, B., Farinas, I., Karsenty, G. and Grosschedl, R. (2006). "SATB2 is a multifunctional determinant of craniofacial patterning and osteoblast differentiation." Cell **125**(5): 971-986.

Dong, M., How, T., Kirkbride, K. C., Gordon, K. J., Lee, J. D., Hempel, N., Kelly, P., Moeller, B. J., Marks, J. R. and Blobe, G. C. (2007). "The type III TGF-beta receptor suppresses breast cancer progression." J Clin Invest **117**(1): 206-217.

Dorsam, R. T. and Gutkind, J. S. (2007). "G-protein-coupled receptors and cancer." Nat Rev Cancer **7**(2): 79-94.

Duan, L., Wu, A. H., Sullivan-Halley, J. and Bernstein, L. (2008). "Nonsteroidal anti-inflammatory drugs and risk of esophageal and gastric adenocarcinomas in Los Angeles County." Cancer Epidemiol Biomarkers Prev **17**(1): 126-134.

Dudbridge, F. (2008). "Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data." Hum Hered **66**(2): 87-98.

Dulak, A. M., Stojanov, P., Peng, S., Lawrence, M. S., Fox, C., Stewart, C., Bandla, S., Imamura, Y., Schumacher, S. E., Shefler, E., McKenna, A., Carter, S. L., Cibulskis, K., Sivachenko, A., Saksena, G., Voet, D., Ramos, A. H., Auclair, D., Thompson, K., Sougnez, C., Onofrio, R. C., Guiducci, C., Beroukhim, R., Zhou, Z., Lin, L., Lin, J., Reddy, R., Chang, A., Landrenau, R., Pennathur, A., Ogino, S., Luketich, J. D., Golub, T. R., Gabriel, S. B., Lander, E. S., Beer, D. G., Godfrey, T. E., Getz, G. and Bass, A. J. (2013). "Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity." Nat Genet **45**(5): 478-486.

Eberhard, J., Gaber, A., Wangefjord, S., Nodin, B., Uhlen, M., Ericson Lindquist, K. and Jirstrom, K. (2012). "A cohort study of the prognostic and

treatment predictive value of SATB2 expression in colorectal cancer." Br J Cancer **106**(5): 931-938.

Ebermann, I., Wiesen, M. H., Zrenner, E., Lopez, I., Pigeon, R., Kohl, S., Lowenheim, H., Koenekoop, R. K. and Bolz, H. J. (2009). "GPR98 mutations cause Usher syndrome type 2 in males." J Med Genet **46**(4): 277-280.

Efeyan, A. and Serrano, M. (2007). "p53: guardian of the genome and policeman of the oncogenes." Cell Cycle **6**(9): 1006-1010.

Egashira, A., Morita, M., Kakeji, Y., Sadanaga, N., Oki, E., Honbo, T., Ohta, M. and Maehara, Y. (2007). "p53 Gene mutations in esophageal squamous cell carcinoma and their relevance to etiology and pathogenesis: Results in Japan and comparisons with other countries." Cancer Science **98**(8): 1152-1156.

ENCODE Project Consortium (2011). "A user's guide to the encyclopedia of DNA elements (ENCODE)." PLoS Biol **9**(4): e1001046.

Esteller, M. and Herman, J. G. (2004). "Generating mutations but providing chemosensitivity: the role of O6-methylguanine DNA methyltransferase in human cancer." Oncogene **23**(1): 1-8.

Esteve, A., Martel-Planche, G., Sylla, B. S., Hollstein, M., Hainaut, P. and Montesano, R. (1996). "Low frequency of p16/CDKN2 gene mutations in esophageal carcinomas." Int J Cancer **66**(3): 301-304.

Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P., Zhernakova, A., Stahl, E., Viatte, S., McAllister, K., Amos, C. I., Padyukov, L., Toes, R. E., Huizinga, T. W., Wijmenga, C., Trynka, G., Franke, L., Westra, H. J., Alfredsson, L., Hu, X., Sandor, C., de Bakker, P. I., Davila, S., Khor, C. C., Heng, K. K., Andrews, R., Edkins, S., Hunt, S. E., Langford, C., Symmons, D., Concannon, P., Onengut-Gumuscu, S., Rich, S. S., Deloukas, P., Gonzalez-Gay, M. A., Rodriguez-Rodriguez, L., Arlsetig, L., Martin, J., Rantapaa-Dahlqvist, S., Plenge, R. M., Raychaudhuri, S., Klareskog, L., Gregersen, P. K. and Worthington, J. (2012). "High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis." Nat Genet **44**(12): 1336-1340.

Farrow, D. C., Vaughan, T. L., Sweeney, C., Gammon, M. D., Chow, W. H., Risch, H. A., Stanford, J. L., Hansten, P. D., Mayne, S. T., Schoenberg, J. B., Rotterdam, H., Ahsan, H., West, A. B., Dubrow, R., Fraumeni, J. F., Jr. and Blot, W. J. (2000). "Gastroesophageal reflux disease, use of H2 receptor antagonists, and risk of esophageal and gastric cancer." Cancer Causes Control **11**(3): 231-238.

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C. and Parkin, D. M. (2010.a). "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008." Int J Cancer **127**(12): 2893-2917.

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C. D. and Parkin, D. M. (2010.b). "GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10 [Internet].".

Finger, E. C., Turley, R. S., Dong, M., How, T., Fields, T. A. and Blobe, G. C. (2008). "TbetaRIII suppresses non-small cell lung cancer invasiveness and tumorigenicity." Carcinogenesis **29**(3): 528-535.

Flaherty, K. T., Puzanov, I., Kim, K. B., Ribas, A., McArthur, G. A., Sosman, J. A., O'Dwyer, P. J., Lee, R. J., Grippo, J. F., Nolop, K. and Chapman, P. B. (2010). "Inhibition of mutated, activated BRAF in metastatic melanoma." N Engl J Med **363**(9): 809-819.

Fletcher, O. and Houlston, R. S. (2010). "Architecture of inherited susceptibility to common cancer." Nat Rev Cancer **10**(5): 353-361.

Folsom, A. R., Peacock, J. M. and Boerwinkle, E. (2007). "Sequence variation in proprotein convertase subtilisin/kexin type 9 serine protease gene, low LDL cholesterol, and cancer incidence." Cancer Epidemiol Biomarkers Prev **16**(11): 2455-2458.

Frank, B., Rigas, S. H., Bermejo, J. L., Wiestler, M., Wagner, K., Hemminki, K., Reed, M. W., Sutter, C., Wappenschmidt, B., Balasubramanian, S. P., Meindl, A., Kiechle, M., Bugert, P., Schmutzler, R. K., Bartram, C. R., Justenhoven, C., Ko, Y. D., Bruning, T., Brauch, H., Hamann, U., Pharoah, P. P., Dunning, A. M., Pooley, K. A., Easton, D. F., Cox, A. and Burwinkel, B. (2008). "The CASP8 -652 6N del promoter polymorphism and breast cancer risk: a multicenter study." Breast Cancer Res Treat **111**(1): 139-144.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., Anderson, C. A., Bis, J. C., Bumpstead, S., Ellinghaus, D., Festen, E. M., Georges, M., Green, T., Haritunians, T., Jostins, L., Latiano, A., Mathew, C. G., Montgomery, G. W., Prescott, N. J., Raychaudhuri, S., Rotter, J. I., Schumm, P., Sharma, Y., Simms, L. A., Taylor, K. D., Whiteman, D., Wijmenga, C., Baldassano, R. N., Barclay, M., Bayless, T. M., Brand, S., Buning, C., Cohen, A., Colombel, J. F., Cottone, M., Stronati, L., Denson, T., De Vos, M., D'Inca, R., Dubinsky, M., Edwards, C., Florin, T., Franchimont, D., Gearry, R., Glas, J., Van Gossum, A., Guthery, S. L., Halfvarson, J., Verspaget, H. W., Hugot, J. P., Karban, A., Laukens, D.,

Lawrance, I., Lemann, M., Levine, A., Libioulle, C., Louis, E., Mowat, C., Newman, W., Panes, J., Phillips, A., Proctor, D. D., Regueiro, M., Russell, R., Rutgeerts, P., Sanderson, J., Sans, M., Seibold, F., Steinhart, A. H., Stokkers, P. C., Torkvist, L., Kullak-Ublick, G., Wilson, D., Walters, T., Targan, S. R., Brant, S. R., Rioux, J. D., D'Amato, M., Weersma, R. K., Kugathasan, S., Griffiths, A. M., Mansfield, J. C., Vermeire, S., Duerr, R. H., Silverberg, M. S., Satsangi, J., Schreiber, S., Cho, J. H., Annese, V., Hakonarson, H., Daly, M. J. and Parkes, M. (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci." Nat Genet **42**(12): 1118-1125.

Friend, S. H., Bernards, R., Rogelj, S., Weinberg, R. A., Rapaport, J. M., Albert, D. M. and Dryja, T. P. (1986). "A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma." Nature **323**(6089): 643-646.

Fuku, T., Semba, S., Yutori, H. and Yokozaki, H. (2007). "Increased wild-type p53-induced phosphatase 1 (Wip1 or PPM1D) expression correlated with downregulation of checkpoint kinase 2 in human gastric carcinoma." Pathol Int **57**(9): 566-571.

Gamieldien, W., Victor, T. C., Mugwanya, D., Stepien, A., Gelderblom, W. C., Marasas, W. F., Geiger, D. H. and van Helden, P. D. (1998). "p53 and p16/CDKN2 gene mutations in esophageal tumors from a high-incidence area in South Africa." Int J Cancer **78**(5): 544-549.

Gammon, M. D., Schoenberg, J. B., Ahsan, H., Risch, H. A., Vaughan, T. L., Chow, W. H., Rotterdam, H., West, A. B., Dubrow, R., Stanford, J. L., Mayne, S. T., Farrow, D. C., Niwa, S., Blot, W. J. and Fraumeni, J. F., Jr. (1997). "Tobacco, alcohol, and socioeconomic status and adenocarcinomas of the esophagus and gastric cardia." J Natl Cancer Inst **89**(17): 1277-1284.

Gao, Y., Hu, N., Han, X., Giffen, C., Ding, T., Goldstein, A. and Taylor, P. (2009). "Family history of cancer and risk for esophageal and gastric cancer in Shanxi, China." BMC Cancer **9**: 269.

Garavello, W., Negri, E., Talamini, R., Levi, F., Zambon, P., Dal Maso, L., Bosetti, C., Franceschi, S. and La Vecchia, C. (2005). "Family history of cancer, its combination with smoking and drinking, and risk of squamous cell carcinoma of the esophagus." Cancer Epidemiol Biomarkers Prev **14**(6): 1390-1393.

Gatza, C. E., Oh, S. Y. and Blobe, G. C. (2010). "Roles for the type III TGF-beta receptor in human cancer." Cell Signal **22**(8): 1163-1174.

Gauderman, W. J. (2002). "Sample size requirements for association studies of gene-gene interaction." Am J Epidemiol **155**(5): 478-484.

Ghadirian, P. (1985). "Familial history of esophageal cancer." Cancer **56**(8): 2112-2116.

Ghavami, S., Hashemi, M., Ande, S. R., Yeganeh, B., Xiao, W., Eshraghi, M., Bus, C. J., Kadkhoda, K., Wiechec, E., Halayko, A. J. and Los, M. (2009). "Apoptosis and cancer: mutations within caspase genes." J Med Genet **46**(8): 497-510.

Giroux, M. A., Audrezet, M. P., Metges, J. P., Lozac'h, P., Volant, A., Nousbaum, J. B., Labat, J. P., Gouerou, H., Ferec, C. and Robaszkiewicz, M. (2002). "Infrequent p16/CDKN2 alterations in squamous cell carcinoma of the oesophagus." Eur J Gastroenterol Hepatol **14**(1): 15-18.

Glick, A. B. (2012). "The Role of TGF Signaling in Squamous Cell Cancer: Lessons from Mouse Models." J Skin Cancer **2012**(Article ID 249063).

Gokhale, N. A., Zaremba, A. and Shears, S. B. (2011). "Receptor-dependent compartmentalization of PPIP5K1, a kinase with a cryptic polyphosphoinositide binding domain." Biochem J **434**(3): 415-426.

Gordon, K. J. and Blobe, G. C. (2008). "Role of transforming growth factor-beta superfamily signaling pathways in human disease." Biochim Biophys Acta **1782**(4): 197-228.

Gordon, K. J., Dong, M., Chislock, E. M., Fields, T. A. and Blobe, G. C. (2008). "Loss of type III transforming growth factor beta receptor expression increases motility and invasiveness associated with epithelial to mesenchymal transition during pancreatic cancer progression." Carcinogenesis **29**(2): 252-262.

Gu, H., Ding, G., Zhang, W., Liu, C., Chen, Y., Chen, S. and Jiang, P. (2012). "Replication study of PLCE1 and C20orf54 polymorphism and risk of esophageal cancer in a Chinese population." Mol Biol Rep **39**(9): 9105-9111.

Gu, J., Ajani, J. A., Hawk, E. T., Ye, Y., Lee, J. H., Bhutani, M. S., Hofstetter, W. L., Swisher, S. G., Wang, K. K. and Wu, X. (2010). "Genome-wide catalogue of chromosomal aberrations in barrett's esophagus and esophageal adenocarcinoma: a high-density single nucleotide polymorphism array analysis." Cancer Prev Res (Phila) **3**(9): 1176-1186.

Gudmundsson, J., Johannesdottir, G., Bergthorsson, J. T., Arason, A., Ingvarsson, S., Egilsson, V. and Barkardottir, R. B. (1995). "Different tumor types from BRCA2 carriers show wild-type chromosome deletions on 13q12q13." Cancer Research **55**(21): 4830-4832.

Guo, C., Chang, C. C., Wortham, M., Chen, L. H., Kernagis, D. N., Qin, X., Cho, Y. W., Chi, J. T., Grant, G. A., McLendon, R. E., Yan, H., Ge, K., Papadopoulos, N., Bigner, D. D. and He, Y. (2012). "Global identification of MLL2-targeted loci reveals MLL2's role in diverse signaling pathways." Proc Natl Acad Sci U S A **109**(43): 17603-17608.

Guo, W., Wang, N., Li, Y. and Zhang, J. H. (2005). "Polymorphisms in tumor necrosis factor genes and susceptibility to esophageal squamous cell carcinoma and gastric cardiac adenocarcinoma in a population of high incidence region of North China." Chin Med J (Engl) **118**(22): 1870-1878.

Haiman, C. A., Garcia, R. R., Kolonel, L. N., Henderson, B. E., Wu, A. H. and Le Marchand, L. (2008). "A promoter polymorphism in the CASP8 gene is not associated with cancer risk." Nat Genet **40**(3): 259-260.

Hale, M. J., Liptz, T. R. and Paterson, A. C. (1989). "Association between human papillomavirus and carcinoma of the esophagus in South-African Blacks. A histochemical and immunohistochemical study." S Afr Med J **76**(7): 329-330.

Hall, J., Hashibe, M., Boffetta, P., Gaborieau, V., Moullan, N., Chabrier, A., Zaridze, D., Shangina, O., Szeszenia-Dabrowska, N., Mates, D., Janout, V., Fabianova, E., Holcatova, I., Hung, R. J., McKay, J., Canzian, F. and Brennan, P. (2007). "The association of sequence variants in DNA repair and cell cycle genes with cancers of the upper aerodigestive tract." Carcinogenesis **28**(3): 665-671.

Hamilton, S. R. and Aaltonen, L. A. (2000). World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Digestive System. Lyon IARC Press.

Hanahan, D. and Weinberg, R. A. (2000). "The hallmarks of cancer." Cell **100**(1): 57-70.

Hanahan, D. and Weinberg, R. A. (2011). "Hallmarks of Cancer: The Next Generation." Cell **144**(5): 646-674.

Hashibe, M., Boffetta, P., Zaridze, D., Shangina, O., Szeszenia-Dabrowska, N., Mates, D., Janout, V., Fabianova, E., Bencko, V., Moullan, N., Chabrier, A., Hung, R., Hall, J., Canzian, F. and Brennan, P. (2006). "Evidence for an important role of alcohol- and aldehyde-metabolizing genes in cancers

of the upper aerodigestive tract." Cancer Epidemiol Biomarkers Prev **15**(4): 696-703.

Hashibe, M., McKay, J. D., Curado, M. P., Oliveira, J. C., Koifman, S., Koifman, R., Zaridze, D., Shangina, O., Wunsch-Filho, V., Eluf-Neto, J., Levi, J. E., Matos, E., Lagiou, P., Lagiou, A., Benhamou, S., Bouchardy, C., Szeszenia-Dabrowska, N., Menezes, A., Dall'Agnol, M. M., Merletti, F., Richiardi, L., Fernandez, L., Lence, J., Talamini, R., Barzan, L., Mates, D., Mates, I. N., Kjaerheim, K., Macfarlane, G. J., Macfarlane, T. V., Simonato, L., Canova, C., Holcatova, I., Agudo, A., Castellsague, X., Lowry, R., Janout, V., Kollarova, H., Conway, D. I., McKinney, P. A., Znaor, A., Fabianova, E., Bencko, V., Lissowska, J., Chabrier, A., Hung, R. J., Gaborieau, V., Boffetta, P. and Brennan, P. (2008). "Multiple ADH genes are associated with upper aerodigestive cancers." Nat Genet **40**(6): 707-709.

He, Y., Ye, L., Shan, B., Song, G., Meng, F. and Wang, S. (2009). "Effect of riboflavin-fortified salt nutrition intervention on esophageal squamous cell carcinoma in a high incidence area, China." Asian Pac J Cancer Prev **10**(4): 619-622.

Hemminki, K. and Jiang, Y. (2002). "Familial and second esophageal cancers: a nation-wide epidemiologic study from Sweden." Int J Cancer **98**(1): 106-109.

Henderson, P., van Limbergen, J. E., Wilson, D. C., Satsangi, J. and Russell, R. K. (2011). "Genetics of childhood-onset inflammatory bowel disease." Inflamm Bowel Dis **17**(1): 346-361.

Hendricks, D. and Parker, M. I. (2002). "Oesophageal cancer in Africa." IUBMB Life **53**(4-5): 263-268.

Henn, B. M., Gignoux, C. R., Jobin, M., Granka, J. M., Macpherson, J. M., Kidd, J. M., Rodriguez-Botigue, L., Ramachandran, S., Hon, L., Brisbin, A., Lin, A. A., Underhill, P. A., Comas, D., Kidd, K. K., Norman, P. J., Parham, P., Bustamante, C. D., Mountain, J. L. and Feldman, M. W. (2011). "Hunter-gatherer genomic diversity suggests a southern African origin for modern humans." Proc Natl Acad Sci U S A **108**(13): 5154-5162.

Hermiston, M. L., Xu, Z. and Weiss, A. (2003). "CD45: a critical regulator of signaling thresholds in immune cells." Annu Rev Immunol **21**: 107-137.

Hindorff, L. A., MacArthur, J., Morales, J., Junkins, H. A., Hall, P. N., Klemm, A. K. and Manolio, T. A. "A Catalog of Published Genome-Wide Association Studies." (Available at: www.genome.gov/gwastudies. Accessed 15/12/2012.).

Hollstein, M. C., Metcalf, R. A., Welsh, J. A., Montesano, R. and Harris, C. C. (1990). "Frequent mutation of the p53 gene in human esophageal cancer." Proc Natl Acad Sci U S A **87**(24): 9958-9961.

Holmes, R. S. and Vaughan, T. L. (2007). "Epidemiology and pathogenesis of esophageal cancer." Semin Radiat Oncol **17**(1): 2-9.

Hu, H., Yang, J., Sun, Y., Yang, Y., Qian, J., Jin, L., Wang, M., Bi, R., Zhang, R., Zhu, M., Sun, M., Ma, H., Wei, Q., Jiang, G., Zhou, X. and Chen, H. (2012). "Putatively Functional PLCE1 Variants and Susceptibility to Esophageal Squamous Cell Carcinoma (ESCC): A Case-Control Study in Eastern Chinese Populations." Ann Surg Oncol **19**(7): 2403-2410.

Hu, N., Wang, C., Hu, Y., Yang, H. H., Kong, L. H., Lu, N., Su, H., Wang, Q. H., Goldstein, A. M., Buetow, K. H., Emmert-Buck, M. R., Taylor, P. R. and Lee, M. P. (2006). "Genome-wide loss of heterozygosity and copy number alteration in esophageal squamous cell carcinoma using the Affymetrix GeneChip Mapping 10 K array." BMC Genomics **7**: 299.

Hu, N., Wang, C., Ng, D., Clifford, R., Yang, H. H., Tang, Z. Z., Wang, Q. H., Han, X. Y., Giffen, C., Goldstein, A. M., Taylor, P. R. and Lee, M. P. (2009). "Genomic characterization of esophageal squamous cell carcinoma from a high-risk population in China." Cancer Res **69**(14): 5908-5917.

Hu, N., Wang, C., Su, H., Li, W. J., Emmert-Buck, M. R., Li, G., Roth, M. J., Tang, Z. Z., Lu, N., Giffen, C., Albert, P. S., Taylor, P. R. and Goldstein, A. M. (2004). "High frequency of CDKN2A alterations in esophageal squamous cell carcinoma from a high-risk Chinese population." Genes Chromosomes Cancer **39**(3): 205-216.

Hu, Z., Wu, C., Shi, Y., Guo, H., Zhao, X., Yin, Z., Yang, L., Dai, J., Hu, L., Tan, W., Li, Z., Deng, Q., Wang, J., Wu, W., Jin, G., Jiang, Y., Yu, D., Zhou, G., Chen, H., Guan, P., Chen, Y., Shu, Y., Xu, L., Liu, X., Liu, L., Xu, P., Han, B., Bai, C., Zhao, Y., Zhang, H., Yan, Y., Ma, H., Chen, J., Chu, M., Lu, F., Zhang, Z., Chen, F., Wang, X., Jin, L., Lu, J., Zhou, B., Lu, D., Wu, T., Lin, D. and Shen, H. (2011). "A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese." Nat Genet **43**(8): 792-796.

Huang, W. Y., Olshan, A. F., Schwartz, S. M., Berndt, S. I., Chen, C., Llaca, V., Chanock, S. J., Fraumeni, J. F., Jr. and Hayes, R. B. (2005). "Selected genetic polymorphisms in MGMT, XRCC1, XPD, and XRCC3 and risk of head and neck cancer: a pooled analysis." Cancer Epidemiol Biomarkers Prev **14**(7): 1747-1753.

IARC (2002). IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Volume 82. Some Traditional Herbal Medicines, Some Mycotoxins, Naphthalene and Styrene. Lyon, France, International Agency for Research on Cancer.

IARC (2003). Cancer in Africa: Epidemiology and prevention Parkin, D. M., Ferlay, J., Hamdi-Cherif, M., Sitas, F., Thomas, J. O., Wabinga, H. and Whelan, S. L. Lyon, Fance, International Agency for Research on Cancer, and World Health Organization.

IARC (2004.a). IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Volume 83. Tobacco Smoke and Involuntary Smoking. Lyon, France, International Agency for Research on Cancer.

IARC (2004.b). IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Volume 85. Betel-quid and areca-nut chewing and some areca-nut derived nitrosamines. Lyon, France, International Agency for Research on Cancer.

IARC (2007.a). Cancer Incidence in Five Continents. Vol IX. Curado, M. P., Edwards, B., Shin, H. R., Storm, H., Ferlay, J., Heanue, M. and Boyle, P. Lyon, France, International Agency for Research on Cancer.

IARC (2007.b). IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Volume 90. Human Papillomaviruses. Lyon, France, International Agency for Research on Cancer.

IARC (2008). World Cancer Report 2008. Boyle, P. and Levin, B. Lyon, France, International Agency for Research on Cancer.

IARC (2010). IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Volume 96. Alcohol Consumption and Ethyl Carbamate. Lyon, France, International Agency for Research on Cancer.

Insel, P. A., Zhang, L., Murray, F., Yokouchi, H. and Zambon, A. C. (2012). "Cyclic AMP is both a pro-apoptotic and anti-apoptotic second messenger." Acta Physiologica **204**(2): 277-287.

Islami, F., Boffetta, P., Ren, J. S., Pedoeim, L., Khatib, D. and Kamangar, F. (2009.b). "High-temperature beverages and foods and esophageal cancer risk--a systematic review." Int J Cancer **125**(3): 491-524.

Islami, F. and Kamangar, F. (2008). "Helicobacter pylori and esophageal cancer risk: a meta-analysis." Cancer Prev Res (Phila) **1**(5): 329-338.

Islami, F., Kamangar, F., Nasrollahzadeh, D., Moller, H., Boffetta, P. and Malekzadeh, R. (2009.a). "Oesophageal cancer in Golestan Province, a high-incidence area in northern Iran - A review." Eur J Cancer **45**(18): 3156-3165.

Islami, F., Pourshams, A., Nasrollahzadeh, D., Kamangar, F., Fahimi, S., Shakeri, R., Abedi-Ardekani, B., Merat, S., Vahedi, H., Semnani, S., Abnet, C. C., Brennan, P., Moller, H., Saidi, F., Dawsey, S. M., Malekzadeh, R. and Boffetta, P. (2009.c). "Tea drinking habits and oesophageal cancer in a high risk area in northern Iran: population based case-control study." BMJ **338**: b929.

Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., Kivinen, K., Bojang, K. A., Conway, D. J., Pinder, M., Sirugo, G., Sisay-Joof, F., Usen, S., Auburn, S., Bumpstead, S. J., Campino, S., Coffey, A., Dunham, A., Fry, A. E., Green, A., Gwilliam, R., Hunt, S. E., Inouye, M., Jeffreys, A. E., Mendy, A., Palotie, A., Potter, S., Ragoussis, J., Rogers, J., Rowlands, K., Somaskantharajah, E., Whittaker, P., Widden, C., Donnelly, P., Howie, B., Marchini, J., Morris, A., SanJoaquin, M., Achidi, E. A., Agbenyega, T., Allen, A., Amodu, O., Corran, P., Djimde, A., Dolo, A., Doumbo, O. K., Drakeley, C., Dunstan, S., Evans, J., Farrar, J., Fernando, D., Hien, T. T., Horstmann, R. D., Ibrahim, M., Karunaweera, N., Kokwaro, G., Koram, K. A., Lemnge, M., Makani, J., Marsh, K., Michon, P., Modiano, D., Molyneux, M. E., Mueller, I., Parker, M., Peshu, N., Plowe, C. V., Puijalon, O., Reeder, J., Reyburn, H., Riley, E. M., Sakuntabhai, A., Singhasivanon, P., Sirima, S., Tall, A., Taylor, T. E., Thera, M., Troye-Blomberg, M., Williams, T. N., Wilson, M. and Kwiatkowski, D. P. (2009). "Genome-wide and fine-resolution association analysis of malaria in West Africa." Nat Genet **41**(6): 657-665.

Jemal, A., Bray, F., Forman, D., O'Brien, M., Ferlay, J., Center, M. and Parkin, D. M. (2012). "Cancer burden in Africa and opportunities for prevention." Cancer **118**(18): 4372-4384.

Ji, A., Wang, J., Yang, J., Wei, Z., Lian, C., Ma, L., Chen, J., Qin, X., Wang, L. and Wei, W. (2011). "Functional SNPs in human C20orf54 gene influence susceptibility to esophageal squamous cell carcinoma." Asian Pac J Cancer Prev **12**(12): 3207-3212.

Johnson, R. W. (2004). South Africa. The first man, the last nation. London, Weidenfeld & Nicolson.

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., Essers, J., Mitrovic, M., Ning, K., Cleynen, I., Theatre, E., Spain, S. L., Raychaudhuri, S., Goyette, P., Wei, Z., Abraham, C., Achkar, J. P.,

Ahmad, T., Amininejad, L., Ananthakrishnan, A. N., Andersen, V., Andrews, J. M., Baidoo, L., Balschun, T., Bampton, P. A., Bitton, A., Boucher, G., Brand, S., Buning, C., Cohain, A., Cichon, S., D'Amato, M., De Jong, D., Devaney, K. L., Dubinsky, M., Edwards, C., Ellinghaus, D., Ferguson, L. R., Franchimont, D., Fransen, K., Gearry, R., Georges, M., Gieger, C., Glas, J., Haritunians, T., Hart, A., Hawkey, C., Hedl, M., Hu, X., Karlsen, T. H., Kupcinskas, L., Kugathasan, S., Latiano, A., Laukens, D., Lawrance, I. C., Lees, C. W., Louis, E., Mahy, G., Mansfield, J., Morgan, A. R., Mowat, C., Newman, W., Palmieri, O., Ponsioen, C. Y., Potocnik, U., Prescott, N. J., Regueiro, M., Rotter, J. I., Russell, R. K., Sanderson, J. D., Sans, M., Satsangi, J., Schreiber, S., Simms, L. A., Sventoraityte, J., Targan, S. R., Taylor, K. D., Tremelling, M., Verspaget, H. W., De Vos, M., Wijmenga, C., Wilson, D. C., Winkelmann, J., Xavier, R. J., Zeissig, S., Zhang, B., Zhang, C. K., Zhao, H., Silverberg, M. S., Annese, V., Hakonarson, H., Brant, S. R., Radford-Smith, G., Mathew, C. G., Rioux, J. D., Schadt, E. E., Daly, M. J., Franke, A., Parkes, M., Vermeire, S., Barrett, J. C. and Cho, J. H. (2012). "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease." Nature **491**(7422): 119-124.

Kaina, B., Christmann, M., Naumann, S. and Roos, W. P. (2007). "MGMT: key node in the battle against genotoxicity, carcinogenicity and apoptosis induced by alkylating agents." DNA Repair (Amst) **6**(8): 1079-1099.

Kaklamani, V. G., Hou, N., Bian, Y., Reich, J., Offit, K., Michel, L. S., Rubinstein, W. S., Rademaker, A. and Pasche, B. (2003). "TGFBR1*6A and cancer risk: a meta-analysis of seven case-control studies." J Clin Oncol **21**(17): 3236-3243.

Kamangar, F., Chow, W. H., Abnet, C. C. and Dawsey, S. M. (2009). "Environmental causes of esophageal cancer." Gastroenterol Clin North Am **38**(1): 27-57.

Khor, C. C., Ramdas, W. D., Vithana, E. N., Cornes, B. K., Sim, X., Tay, W. T., Saw, S. M., Zheng, Y., Lavanya, R., Wu, R., Wang, J. J., Mitchell, P., Uitterlinden, A. G., Rivadeneira, F., Teo, Y. Y., Chia, K. S., Seielstad, M., Hibberd, M., Vingerling, J. R., Klaver, C. C., Jansonius, N. M., Tai, E. S., Wong, T. Y., van Duijn, C. M. and Aung, T. (2011). "Genome-wide association studies in Asians confirm the involvement of ATOH7 and TGFBR3, and further identify CARD10 as a novel locus influencing optic disc area." Hum Mol Genet **20**(9): 1864-1872.

Kitada, S., Andersen, J., Akar, S., Zapata, J. M., Takayama, S., Krajewski, S., Wang, H. G., Zhang, X., Bullrich, F., Croce, C. M., Rai, K., Hines, J. and Reed, J. C. (1998). "Expression of apoptosis-regulating proteins in

chronic lymphocytic leukemia: correlations with In vitro and In vivo chemoresponses." Blood **91**(9): 3379-3389.

Knudson, A. G., Jr. (1971). "Mutation and cancer: statistical study of retinoblastoma." Proc Natl Acad Sci U S A **68**(4): 820-823.

Kozarewa, I., Rosa-Rosa, J. M., Wardell, C. P., Walker, B. A., Fenwick, K., Assiotis, I., Mitsopoulos, C., Zvelebil, M., Morgan, G. J., Ashworth, A. and Lord, C. J. (2012). "A modified method for whole exome resequencing from minimal amounts of starting DNA." PLoS One **7**(3): e32617.

Kritchevsky, S. B., Wilcosky, T. C., Morris, D. L., Truong, K. N. and Tyroler, H. A. (1991). "Changes in plasma lipid and lipoprotein cholesterol and weight prior to the diagnosis of cancer." Cancer Res **51**(12): 3198-3203.

Kuiper, R. P., Ligtenberg, M. J., Hoogerbrugge, N. and Geurts van Kessel, A. (2010). "Germline copy number variation and cancer risk." Curr Opin Genet Dev **20**(3): 282-289.

Kuo, B. and Urma, D. (2006). "Esophagus - anatomy and development." GI Motility online **Available from http://www.nature.com/gimo/contents/pt1/full/gimo6.html**.

Kurose, K., Mine, N., Doi, D., Ota, Y., Yoneyama, K., Konishi, H., Araki, T. and Emi, M. (2000). "Novel gene fusion of COX6C at 8q22-23 to HMGIC at 12q15 in a uterine leiomyoma." Genes Chromosomes Cancer **27**(3): 303-307.

Lachance, J., Vernot, B., Elbers, C. C., Ferwerda, B., Froment, A., Bodo, J. M., Lema, G., Fu, W., Nyambo, T. B., Rebbeck, T. R., Zhang, K., Akey, J. M. and Tishkoff, S. A. (2012). "Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse african hunter-gatherers." Cell **150**(3): 457-469.

Lagergren, J., Bergstrom, R., Lindgren, A. and Nyren, O. (1999). "Symptomatic gastroesophageal reflux as a risk factor for esophageal adenocarcinoma." N Engl J Med **340**(11): 825-831.

Lagergren, J., Bergstrom, R., Lindgren, A. and Nyren, O. (2000). "The role of tobacco, snuff and alcohol use in the aetiology of cancer of the oesophagus and gastric cardia." Int J Cancer **85**(3): 340-346.

Lam, A. K. (2000). "Molecular biology of esophageal squamous cell carcinoma." Crit Rev Oncol Hematol **33**(2): 71-90.

Landemaine, T., Jackson, A., Bellahcene, A., Rucci, N., Sin, S., Abad, B. M., Sierra, A., Boudinet, A., Guinebretiere, J. M., Ricevuto, E., Nogues, C., Briffod, M., Bieche, I., Cherel, P., Garcia, T., Castronovo, V., Teti, A., Lidereau, R. and Driouch, K. (2008). "A six-gene signature predicting breast cancer lung metastasis." Cancer Res **68**(15): 6092-6099.

Lao-Sirieix, P., Caldas, C. and Fitzgerald, R. C. (2010). "Genetic predisposition to gastro-oesophageal cancer." Curr Opin Genet Dev **20**(3): 210-217.

Law, M. R. and Thompson, S. G. (1991). "Low serum-cholesterol and the risk of cancer: an analysis of the published prospective studies." Cancer Causes Control **2**(4): 253-261.

Layke, J. C. and Lopez, P. P. (2006). "Esophageal cancer: a review and update." Am Fam Physician **73**(12): 2187-2194.

Lazarevic, V., Chen, X., Shim, J. H., Hwang, E. S., Jang, E., Bolm, A. N., Oukka, M., Kuchroo, V. K. and Glimcher, L. H. (2011). "T-bet represses T(H)17 differentiation by preventing Runx1-mediated activation of the gene encoding RORgammat." Nat Immunol **12**(1): 96-104.

Lazarini, M., Traina, F., Machado-Neto, J. A., Barcellos, K. S. A., Moreira, Y. B., Brandao, M. M., Verjovski-Almeida, S., Ridley, A. J. and Saad, S. T. O. (2013). "ARHGAP21 is a RhoGAP for RhoA and RhoC with a role in proliferation and migration of prostate adenocarcinoma cells." Biochim Biophys Acta. **1832**(2): 365-374.

Lee, C. H., Lee, J. M., Wu, D. C., Hsu, H. K., Kao, E. L., Huang, H. L., Wang, T. N., Huang, M. C. and Wu, M. T. (2005). "Independent and combined effects of alcohol intake, tobacco smoking and betel quid chewing on the risk of esophageal cancer in Taiwan." Int J Cancer **113**(3): 475-482.

Lee, C. H., Wu, D. C., Wu, I. C., Goan, Y. G., Lee, J. M., Chou, S. H., Chan, T. F., Huang, H. L., Hung, Y. H., Huang, M. C., Lai, T. C., Wang, T. N., Lan, C. C., Tsai, S., Lin, W. Y. and Wu, M. T. (2009). "Genetic modulation of ADH1B and ALDH2 polymorphisms with regard to alcohol and tobacco consumption for younger aged esophageal squamous cell carcinoma diagnosis." Int J Cancer **125**(5): 1134-1142.

Lee, K. H. and Kim, J. R. (2012). "Regulation of HGF-mediated cell proliferation and invasion through NF-kappa B, JunB, and MMP-9 cascades in stomach cancer cells." Clin Exp Metastasis **29**(3): 263-272.

Lettre, G., Palmer, C. D., Young, T., Ejebe, K. G., Allayee, H., Benjamin, E. J., Bennett, F., Bowden, D. W., Chakravarti, A., Dreisbach, A., Farlow, D. N., Folsom, A. R., Fornage, M., Forrester, T., Fox, E., Haiman, C. A.,

Hartiala, J., Harris, T. B., Hazen, S. L., Heckbert, S. R., Henderson, B. E., Hirschhorn, J. N., Keating, B. J., Kritchevsky, S. B., Larkin, E., Li, M., Rudock, M. E., McKenzie, C. A., Meigs, J. B., Meng, Y. A., Mosley, T. H., Newman, A. B., Newton-Cheh, C. H., Paltoo, D. N., Papanicolaou, G. J., Patterson, N., Post, W. S., Psaty, B. M., Qasim, A. N., Qu, L., Rader, D. J., Redline, S., Reilly, M. P., Reiner, A. P., Rich, S. S., Rotter, J. I., Liu, Y., Shrader, P., Siscovick, D. S., Tang, W. H., Taylor, H. A., Tracy, R. P., Vasan, R. S., Waters, K. M., Wilks, R., Wilson, J. G., Fabsitz, R. R., Gabriel, S. B., Kathiresan, S. and Boerwinkle, E. (2011). "Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project." PLoS Genet **7**(2): e1001300.

Lewin, K. J. and Appelman, H. D. (1996). Tumors of the esophagus and stomach. Atlas of tumour pathology. Washington DC, American Registry of Pathology.

Lewis, S. J. and Smith, G. D. (2005). "Alcohol, ALDH2, and esophageal cancer: a meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach." Cancer Epidemiol Biomarkers Prev **14**(8): 1967-1971.

Lewontin, R. C. (1964). "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models." Genetics **49**(1): 49-67.

Li, C., Lu, J., Liu, Z., Wang, L. E., Zhao, H., El-Naggar, A. K., Sturgis, E. M. and Wei, Q. (2010.b). "The six-nucleotide deletion/insertion variant in the CASP8 promoter region is inversely associated with risk of squamous cell carcinoma of the head and neck." Cancer Prev Res (Phila Pa) **3**(2): 246-253.

Li, D., Dandara, C. and Parker, M. I. (2005). "Association of cytochrome P450 2E1 genetic polymorphisms with squamous cell carcinoma of the oesophagus." Clin Chem Lab Med **43**(4): 370-375.

Li, D., Dandara, C. and Parker, M. I. (2010.a). "The 341C/T polymorphism in the GSTP1 gene is associated with increased risk of oesophageal cancer." BMC Genet **11**: 47.

Li, D. P., Dandara, C., Walther, G. and Parker, M. I. (2008). "Genetic polymorphisms of alcohol metabolising enzymes: their role in susceptibility to oesophageal cancer." Clin Chem Lab Med **46**(3): 323-328.

Li, H., Borinskaya, S., Yoshimura, K., Kal'ina, N., Marusin, A., Stepanov, V. A., Qin, Z., Khaliq, S., Lee, M. Y., Yang, Y., Mohyuddin, A., Gurwitz, D.,

Mehdi, S. Q., Rogaev, E., Jin, L., Yankovsky, N. K., Kidd, J. R. and Kidd, K. K. (2009.a). "Refined geographic distribution of the oriental ALDH2*504Lys (nee 487Lys) variant." Ann Hum Genet **73**(3): 335-345.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Proc (2009.b). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, J., Yang, Y., Peng, Y., Austin, R. J., van Eyndhoven, W. G., Nguyen, K. C., Gabriele, T., McCurrach, M. E., Marks, J. R., Hoey, T., Lowe, S. W. and Powers, S. (2002). "Oncogenic properties of PPM1D located within a breast cancer amplification epicenter at 17q23." Nat Genet **31**(2): 133-134.

Li, K. and Yu, P. (2003). "Food groups and risk of esophageal cancer in Chaoshan region of China: a high-risk area of esophageal cancer." Cancer Invest **21**(2): 237-240.

Li, W. J., Hu, N., Su, H., Wang, C., Goldstein, A. M., Wang, Y., Emmert-Buck, M. R., Roth, M. J., Guo, W. J. and Taylor, P. R. (2003). "Allelic loss on chromosome 13q14 and mutation in deleted in cancer 1 gene in esophageal squamous cell carcinoma." Oncogene **22**(2): 314-318.

Li, Y., Schrodi, S., Rowland, C., Tacey, K., Catanese, J. and Grupe, A. (2006). "Genetic evidence for ubiquitin-specific proteases USP24 and USP40 as candidate genes for late-onset Parkinson disease." Hum Mutat **27**(10): 1017-1023.

Lievre, A., Bachet, J. B., Le Corre, D., Boige, V., Landi, B., Emile, J. F., Cote, J. F., Tomasic, G., Penna, C., Ducreux, M., Rougier, P., Penault-Llorca, F. and Laurent-Puig, P. (2006). "KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer." Cancer Res **66**(8): 3992-3995.

Lin, Y., Nie, Y., Zhao, J., Chen, X., Ye, M., Li, Y., Du, Y., Cao, J., Shen, B. and Li, Y. (2012). "Genetic polymorphism at miR-181a binding site contributes to gastric cancer susceptibility." Carcinogenesis **133**(12): 2377–2383.

Lingappa, J. R., Petrovski, S., Kahle, E., Fellay, J., Shianna, K., McElrath, M. J., Thomas, K. K., Baeten, J. M., Celum, C., Wald, A., de Bruyn, G., Mullins, J. I., Nakku-Joloba, E., Farquhar, C., Essex, M., Donnell, D., Kiarie, J., Haynes, B. and Goldstein, D. (2011). "Genomewide association study for determinants of HIV-1 acquisition and viral set point in HIV-1 serodiscordant couples with quantified virus exposure." PLoS One **6**(12): e28632.

Lingwood, R. J., Boyle, P., Milburn, A., Ngoma, T., Arbuthnott, J., McCaffrey, R., Kerr, S. H. and Kerr, D. J. (2008). "The challenge of cancer control in Africa." Nat Rev Cancer **8**(5): 398-403.

Liu, C. Y., Wu, M. C., Chen, F., Ter-Minassian, M., Asomaning, K., Zhai, R., Wang, Z., Su, L., Heist, R. S., Kulke, M. H., Lin, X., Liu, G. and Christiani, D. C. (2010). "A Large-scale genetic association study of esophageal adenocarcinoma risk." Carcinogenesis **31**(7): 1259-1263.

Liu, Y. Z., Wilson, S. G., Wang, L., Liu, X. G., Guo, Y. F., Li, J., Yan, H., Deloukas, P., Soranzo, N., Chinappen-Horsley, U., Cervino, A., Williams, F. M., Xiong, D. H., Zhang, Y. P., Jin, T. B., Levy, S., Papasian, C. J., Drees, B. M., Hamilton, J. J., Recker, R. R., Spector, T. D. and Deng, H. W. (2008). "Identification of PLCL1 gene for hip bone size variation in females in a genome-wide association study." PLoS One **3**(9): e3160.

Lo, H. S., Hu, N., Gere, S., Lu, N., Su, H., Goldstein, A. M., Taylor, P. R. and Lee, M. P. (2002). "Identification of somatic mutations of the RNF6 gene in human esophageal squamous cell carcinoma." Cancer Res **62**(15): 4191-4193.

Lopez, D. (2008). "PCSK9: an enigmatic protease." Biochim Biophys Acta **1781**(4): 184-191.

Lu, H., Ouyang, W. and Huang, C. (2006). "Inflammation, a key event in cancer development." Mol Cancer Res **4**(4): 221-233.

Lu, X., Nguyen, T. A. and Donehower, L. A. (2005). "Reversal of the ATM/ATR-mediated DNA damage response by the oncogenic phosphatase PPM1D." Cell Cycle **4**(8): 1060-1064.

Lu, X., Nguyen, T. A., Moon, S. H., Darlington, Y., Sommer, M. and Donehower, L. A. (2008). "The type 2C phosphatase Wip1: an oncogenic regulator of tumor suppressor and DNA damage response pathways." Cancer Metastasis Rev **27**(2): 123-135.

MacPherson, G., Healey, C. S., Teare, M. D., Balasubramanian, S. P., Reed, M. W., Pharoah, P. D., Ponder, B. A., Meuth, M., Bhattacharyya, N. P. and Cox, A. (2004). "Association of a common variant of the CASP8 gene with reduced risk of breast cancer." J Natl Cancer Inst **96**(24): 1866-1869.

Maelfait, J. and Beyaert, R. (2008). "Non-apoptotic functions of caspase-8." Biochem Pharmacol **76**(11): 1365-1373.

Maeng, C. H., Lee, J., van Hummelen, P., Park, S. H., Palescandolo, E., Jang, J., Park, H. Y., Kang, S. Y., MacConaill, L., Kim, K. M. and Shim, Y. M. (2012). "High-throughput genotyping in metastatic esophageal squamous cell carcinoma identifies phosphoinositide-3-kinase and BRAF mutations." PLoS One **7**(8): e41655.

Magnusson, K., de Wit, M., Brennan, D. J., Johnson, L. B., McGee, S. F., Lundberg, E., Naicker, K., Klinger, R., Kampf, C., Asplund, A., Wester, K., Gry, M., Bjartell, A., Gallagher, W. M., Rexhepaj, E., Kilpinen, S., Kallioniemi, O. P., Belt, E., Goos, J., Meijer, G., Birgisson, H., Glimelius, B., Borrebaeck, C. A., Navani, S., Uhlen, M., O'Connor, D. P., Jirstrom, K. and Ponten, F. (2011). "SATB2 in combination with cytokeratin 20 identifies over 95% of all colorectal carcinomas." Am J Surg Pathol **35**(7): 937-948.

Mahboubi, E., Kmet, J., Cook, P. J., Day, N. E., Ghadirian, P. and Salmasizadeh, S. (1973). "Oesophageal cancer studies in the Caspian Littoral of Iran: the Caspian cancer registry." Br J Cancer **28**(3): 197-214.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. and Visscher, P. M. (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

Marasas, W. F. O., Jaskiewicz, K., Venter, F. S. and Vanschalkwyk, D. J. (1988). "Fusarium moniliforme contamination of maize in esophageal cancer areas in Transkei." S Afr Med J **74**(3): 110-114.

Marasas, W. F. O., Wehner, F. C., Vanrensburg, S. J. and Vanschalkwyk, D. J. (1981). "Mycoflora of Corn Produced in Human Esophageal Cancer Areas in Transkei, Southern-Africa." Phytopathology **71**(8): 792-796.

Massague, J. (2008). "TGFbeta in Cancer." Cell **134**(2): 215-230.

Matejcic, M., Li, D., Prescott, N. J., Lewis, C. M., Mathew, C. G. and Parker, M. I. (2011). "Association of a deletion of GSTT2B with an altered risk of oesophageal squamous cell carcinoma in a South African population: a case-control study." PLoS One **6**(12): e29366.

Mathew, R., Karantza-Wadsworth, V. and White, E. (2007). "Role of autophagy in cancer." Nat Rev Cancer **7**(12): 961-967.

Matsha, T., Brink, L., van Rensburg, S., Hon, D., Lombard, C. and Erasmus, R. (2006). "Traditional home-brewed beer consumption and iron status in patients with esophageal cancer and healthy control subjects from Transkei, South Africa." Nutr Cancer **56**(1): 67-73.

Matsha, T., Erasmus, R., Kafuko, A. B., Mugwanya, D., Stepien, A. and Parker, M. I. (2002). "Human papillomavirus associated with oesophageal cancer." J Clin Pathol **55**(8): 587-590.

Matsha, T. E., Masconi, K., Yako, Y. Y., Hassan, M. S., Macharia, M., Erasmus, R. T. and Kengne, A. P. (2012). "Polymorphisms in the non-muscle Myosin heavy chain gene (MYH9) are associated with lower glomerular filtration rate in mixed ancestry diabetic subjects from South Africa." PLoS One **7**(12): e52529.

McClellan, J. and King, M. C. (2010). "Genetic heterogeneity in human disease." Cell **141**(2): 210-217.

McKay, J. D., Truong, T., Gaborieau, V., Chabrier, A., Chuang, S. C., Byrnes, G., Zaridze, D., Shangina, O., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., Fabianova, E., Bucur, A., Bencko, V., Holcatova, I., Janout, V., Foretova, L., Lagiou, P., Trichopoulos, D., Benhamou, S., Bouchardy, C., Ahrens, W., Merletti, F., Richiardi, L., Talamini, R., Barzan, L., Kjaerheim, K., Macfarlane, G. J., Macfarlane, T. V., Simonato, L., Canova, C., Agudo, A., Castellsague, X., Lowry, R., Conway, D. I., McKinney, P. A., Healy, C. M., Toner, M. E., Znaor, A., Curado, M. P., Koifman, S., Menezes, A., Wunsch-Filho, V., Neto, J. E., Garrote, L. F., Boccia, S., Cadoni, G., Arzani, D., Olshan, A. F., Weissler, M. C., Funkhouser, W. K., Luo, J., Lubinski, J., Trubicka, J., Lener, M., Oszutowska, D., Schwartz, S. M., Chen, C., Fish, S., Doody, D. R., Muscat, J. E., Lazarus, P., Gallagher, C. J., Chang, S. C., Zhang, Z. F., Wei, Q., Sturgis, E. M., Wang, L. E., Franceschi, S., Herrero, R., Kelsey, K. T., McClean, M. D., Marsit, C. J., Nelson, H. H., Romkes, M., Buch, S., Nukui, T., Zhong, S., Lacko, M., Manni, J. J., Peters, W. H., Hung, R. J., McLaughlin, J., Vatten, L., Njolstad, I., Goodman, G. E., Field, J. K., Liloglou, T., Vineis, P., Clavel-Chapelon, F., Palli, D., Tumino, R., Krogh, V., Panico, S., Gonzalez, C. A., Quiros, J. R., Martinez, C., Navarro, C., Ardanaz, E., Larranaga, N., Khaw, K. T., Key, T., Bueno-de-Mesquita, H. B., Peeters, P. H., Trichopoulou, A., Linseisen, J., Boeing, H., Hallmans, G., Overvad, K., Tjonneland, A., Kumle, M., Riboli, E., Valk, K., Voodern, T., Metspalu, A., Zelenika, D., Boland, A., Delepine, M., Foglio, M., Lechner, D., Blanche, H., Gut, I. G., Galan, P., Heath, S., Hashibe, M., Hayes, R. B., Boffetta, P., Lathrop, M. and Brennan, P. (2011). "A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium." PLoS Genet **7**(3): e1001333.

Melhado, R. E. A., A., Tucker, O (2010). "The Changing Face of Esophageal Cancer." Cancers **2**(3): 1379-1404.

Morgan, N. V., Essop, F., Demuth, I., de Ravel, T., Jansen, S., Tischkowitz, M., Lewis, C. M., Wainwright, L., Poole, J., Joenje, H., Digweed, M., Krause, A. and Mathew, C. G. (2005). "A common Fanconi anemia mutation in black populations of sub-Saharan Africa." Blood **105**(9): 3542-3544.

Mori, T., Miura, K., Aoki, T., Nishihira, T., Mori, S. and Nakamura, Y. (1994). "Frequent somatic mutation of the MTS1/CDK4I (multiple tumor suppressor/cyclin-dependent kinase 4 inhibitor) gene in esophageal squamous cell carcinoma." Cancer Res **54**(13): 3396-3397.

Moss, S. F. and Blaser, M. J. (2005). "Mechanisms of disease: Inflammation and the origins of cancer." Nat Clin Pract Oncol **2**(2): 90-97.

Mudan, S. S. and Kang, J.-Y. (2007). Epidemiology and Clinical Presentation in Esophageal Cancer Carcinoma of the Esophagus Rankin, S. C. Cambridge, UK, Cambridge University Press

Mukherji, M., Brill, L. M., Ficarro, S. B., Hampton, G. M. and Schultz, P. G. (2006). "A phosphoproteomic analysis of the ErbB2 receptor tyrosine kinase signaling pathways." Biochemistry **45**(51): 15529-15540.

Nakamura, T., Takeuchi, K., Muraoka, S., Takezoe, H., Takahashi, N. and Mori, N. (1999). "A neurally enriched coronin-like protein, ClipinC, is a novel candidate for an actin cytoskeleton-cortical membrane-linking protein." J Biol Chem **274**(19): 13322-13327.

Nalbant, P., Chang, Y. C., Birkenfeld, J., Chang, Z. F. and Bokoch, G. M. (2009). "Guanine Nucleotide Exchange Factor-H1 Regulates Cell Migration via Localized Activation of RhoA at the Leading Edge." Mol Biol Cell **20**(18): 4070-4082.

Nancarrow, D. J., Handoko, H. Y., Smithers, B. M., Gotley, D. C., Drew, P. A., Watson, D. I., Clouston, A. D., Hayward, N. K. and Whiteman, D. C. (2008). "Genome-wide copy number analysis in esophageal adenocarcinoma using high-density single-nucleotide polymorphism arrays." Cancer Res **68**(11): 4163-4172.

National Cancer Institute (2012). Cancer Trends Progress Report – 2011/2012 Update, National Institute of Health, Bethesda, MD. http://progressreport.cancer.gov.

Ntzani, E. E., Liberopoulos, G., Manolio, T. A. and Ioannidis, J. P. (2012). "Consistency of genome-wide associations across major ancestral groups." Hum Genet **131**(7): 1057-1071.

Okuda, T., van Deursen, J., Hiebert, S. W., Grosveld, G. and Downing, J. R. (1996). "AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis." Cell **84**(2): 321-330.

Olivier, M., Eeles, R., Hollstein, M., Khan, M. A., Harris, C. C. and Hainaut, P. (2002). "The IARC TP53 database: new online mutation analysis and recommendations to users." Hum Mutat **19**(6): 607-614.

Oota, H., Pakstis, A. J., Bonne-Tamir, B., Goldman, D., Grigorenko, E., Kajuna, S. L., Karoma, N. J., Kungulilo, S., Lu, R. B., Odunsi, K., Okonofua, F., Zhukova, O. V., Kidd, J. R. and Kidd, K. K. (2004). "The evolution and population genetics of the ALDH2 locus: random genetic drift, selection, and low levels of recombination." Ann Hum Genet **68**(Pt 2): 93-109.

Oppelt, A., Lobert, V. H., Haglund, K., Mackey, A. M., Rameh, L. E., Liestol, K., Oliver Schink, K., Marie Pedersen, N., Wenzel, E. M., Haugsten, E. M., Brech, A., Erik Rusten, T., Stenmark, H. and Wesche, J. (2012). "Production of phosphatidylinositol 5-phosphate via PIKfyve and MTMR3 regulates cell migration." EMBO Rep **14**(1): 57-64.

Oudes, A. J., Roach, J. C., Walashek, L. S., Eichner, L. J., True, L. D., Vessella, R. L. and Liu, A. Y. (2005). "Application of Affymetrix array and Massively Parallel Signature Sequencing for identification of genes involved in prostate cancer progression." BMC Cancer **5**: 86.

Pacella-Norman, R., Urban, M. I., Sitas, F., Carrara, H., Sur, R., Hale, M., Ruff, P., Patel, M., Newton, R., Bull, D. and Beral, V. (2002). "Risk factors for oesophageal, lung, oral and laryngeal cancers in black South Africans." Br J Cancer **86**(11): 1751-1756.

Palmer, A. J., Lochhead, P., Hold, G. L., Rabkin, C. S., Vaughan, T. L., Lissowska, J., Chow, W. H., Berry, S. and El-Omar, E. M. (2012). "Genetic variation in C20orf54, PLCE1 and MUC1 and the risk of upper gastrointestinal cancers in Caucasian populations." Eur J Cancer Prev **21**(6): 541-544.

Papafili, A., Hill, M. R., Brull, D. J., McAnulty, R. J., Marshall, R. P., Humphries, S. E. and Laurent, G. J. (2002). "Common promoter variant in cyclooxygenase-2 represses gene expression: evidence of role in acute-phase inflammatory response." Arterioscler Thromb Vasc Biol **22**(10): 1631-1636.

Parkin, D. M., Bray, F., Ferlay, J. and Pisani, P. (2005). "Global cancer statistics, 2002." CA Cancer J Clin **55**(2): 74-108.

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B. M., Daly, M. J., Sklar, P., Sullivan, P. F., Bergen, S., Moran, J. L., Hultman, C. M., Lichtenstein, P., Magnusson, P., Purcell, S. M., Haas, D. W., Liang, L., Sunyaev, S., Patterson, N., de Bakker, P. I., Reich, D. and Price, A. L. (2012). "Extremely low-coverage sequencing and imputation increases power for genome-wide association studies." Nat Genet **44**(6): 631-635.

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., O'Brien, S. J., Altshuler, D., Daly, M. J. and Reich, D. (2004). "Methods for high-density admixture mapping of disease genes." Am J Hum Genet **74**(5): 979-1000.

Patterson, N., Petersen, D. C., van der Ross, R. E., Sudoyo, H., Glashoff, R. H., Marzuki, S., Reich, D. and Hayes, V. M. (2010). "Genetic structure of a unique admixed population: implications for medical research." Hum Mol Genet **19**(3): 411-419.

Pera, M., Manterola, C., Vidal, O. and Grande, L. (2005). "Epidemiology of esophageal adenocarcinoma." J Surg Oncol **92**(3): 151-159.

Pereg, Y., Liu, B. Y., O'Rourke, K. M., Sagolla, M., Dey, A., Komuves, L., French, D. M. and Dixit, V. M. (2010). "Ubiquitin hydrolase Dub3 promotes oncogenic transformation by stabilizing Cdc25A." Nat Cell Biol **12**(4): 400-406.

Petitjean, A., Mathe, E., Kato, S., Ishioka, C., Tavtigian, S. V., Hainaut, P. and Olivier, M. (2007). "Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database." Hum Mutat **28**(6): 622-629.

Phillips, W. A., Russell, S. E., Ciavarella, M. L., Choong, D. Y., Montgomery, K. G., Smith, K., Pearson, R. B., Thomas, R. J. and Campbell, I. G. (2006). "Mutation analysis of PIK3CA and PIK3CB in esophageal cancer and Barrett's esophagus." Int J Cancer **118**(10): 2644-2646.

Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Guldemann, T., Kure, B., Mpoloka, S. W., Nakagawa, H., Naumann, C., Lipson, M., Loh, P. R., Lachance, J., Mountain, J., Bustamante, C. D., Berger, B., Tishkoff, S. A., Henn, B. M., Stoneking, M., Reich, D. and Pakendorf, B. (2012). "The genetic prehistory of southern Africa." Nat Commun **3**: 1143.

Planaguma, J., Diaz-Fuertes, M., Gil-Moreno, A., Abal, M., Monge, M., Garcia, A., Baro, T., Thomson, T. M., Xercavins, J., Alameda, F. and Reventos, J. (2004). "A differential gene expression profile reveals overexpression of RUNX1/AML1 in invasive endometrioid carcinoma." Cancer Res **64**(24): 8846-8853.

Pohl, H., Sirovich, B. and Welch, H. G. (2010). "Esophageal adenocarcinoma incidence: are we reaching the peak?" Cancer Epidemiol Biomarkers Prev **19**(6): 1468-1470.

Porcu, M., Kleppe, M., Gianfelici, V., Geerdens, E., De Keersmaecker, K., Tartaglia, M., Foa, R., Soulier, J., Cauwelier, B., Uyttebroeck, A., Macintyre, E., Vandenberghe, P., Asnafi, V. and Cools, J. (2012). "Mutation of the receptor tyrosine phosphatase PTPRC (CD45) in T-cell acute lymphoblastic leukemia." Blood **119**(19): 4476-4479.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat Genet **38**(8): 904-909.

Prickett, T. D., Wei, X., Cardenas-Navia, I., Teer, J. K., Lin, J. C., Walia, V., Gartner, J., Jiang, J., Cherukuri, P. F., Molinolo, A., Davies, M. A., Gershenwald, J. E., Stemke-Hale, K., Rosenberg, S. A., Margulies, E. H. and Samuels, Y. (2011). "Exon capture analysis of G protein-coupled receptors identifies activating mutations in GRM3 in melanoma." Nat Genet **43**(11): 1119-1126.

Pullamsetti, S. S., Banat, G. A., Schmall, A., Szibor, M., Pomagruk, D., Hanze, J., Kolosionek, E., Wilhelm, J., Braun, T., Grimminger, F., Seeger, W., Schermuly, R. T. and Savai, R. (2012). "Phosphodiesterase-4 promotes proliferation and angiogenesis of lung cancer by crosstalk with HIF." Oncogene **32**(9): 1121-1134.

Qiu, L. X., Shi, J., Yuan, H., Jiang, X., Xue, K., Pan, H. F., Li, J. and Zheng, M. H. (2009). "FAS -1,377 G/A polymorphism is associated with cancer susceptibility: evidence from 10,564 cases and 12,075 controls." Hum Genet **125**(4): 431-435.

Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Zaporozhchenko, V., Bormans, C., van Helden, P. D., Hoal, E. G. and Behar, D. M. (2010). "Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture." Am J Hum Genet **86**(4): 611-620.

Ragnarsson, G., Eiriksdottir, G., Johannsdottir, J. T., Jonasson, J. G., Egilsson, V. and Ingvarsson, S. (1999). "Loss of heterozygosity at chromosome 1p

in different solid human tumours: association with survival." Br J Cancer **79**(9-10): 1468-1474.

Ramsay, R. G., Barton, A. L. and Gonda, T. J. (2003). "Targeting c-Myb expression in human disease." Expert Opin Ther Targets **7**(2): 235-248.

Ratnasinghe, D., Tangrea, J., Roth, M. J., Dawsey, S., Hu, N., Anver, M., Wang, Q. H. and Taylor, P. R. (1999). "Expression of cyclooxygenase-2 in human squamous cell carcinoma of the esophagus; an immunohistochemical survey." Anticancer Res **19**(1A): 171-174.

Rebbeck, T. R. (1997). "Molecular epidemiology of the human glutathione S-transferase genotypes GSTM1 and GSTT1 in cancer susceptibility." Cancer Epidemiol Biomarkers Prev **6**(9): 733-743.

Reid, B. J., Li, X., Galipeau, P. C. and Vaughan, T. L. (2010). "Barrett's oesophagus and oesophageal adenocarcinoma: time for a new synthesis." Nat Rev Cancer **10**(2): 87-101.

Reiner, A. P., Ziv, E., Lind, D. L., Nievergelt, C. M., Schork, N. J., Cummings, S. R., Phong, A., Burchard, E. G., Harris, T. B., Psaty, B. M. and Kwok, P. Y. (2005). "Population structure, admixture, and aging-related phenotypes in African American adults: the Cardiovascular Health Study." Am J Hum Genet **76**(3): 463-477.

Rheeder, J. P., Marasas, W. F. O., Thiel, P. G., Sydenham, E. W., Shephard, G. S. and Vanschalkwyk, D. J. (1992). "Fusarium moniliforme and fumonisins in corn in relation to human esophageal cancer in Transkei." Phytopathology **82**(3): 353-357.

Rokkas, T., Pistiolas, D., Sechopoulos, P., Robotis, I. and Margantinis, G. (2007). "Relationship between Helicobacter pylori infection and esophageal neoplasia: a meta-analysis." Clin Gastroenterol Hepatol **5**(12): 1413-1417.

Ronkainen, J., Aro, P., Storskrubb, T., Johansson, S. E., Lind, T., Bolling-Sternevald, E., Vieth, M., Stolte, M., Talley, N. J. and Agreus, L. (2005). "Prevalence of Barrett's esophagus in the general population: an endoscopic study." Gastroenterology **129**(6): 1825-1831.

Roumier, C., Fenaux, P., Lafage, M., Imbert, M., Eclache, V. and Preudhomme, C. (2003). "New mechanisms of AML1 gene alteration in hematological malignancies." Leukemia **17**(1): 9-16.

Ruark, E., Snape, K., Humburg, P., Loveday, C., Bajrami, I., Brough, R., Rodrigues, D. N., Renwick, A., Seal, S., Ramsay, E., Duarte Sdel, V.,

Rivas, M. A., Warren-Perry, M., Zachariou, A., Campion-Flora, A., Hanks, S., Murray, A., Pour, N. A., Douglas, J., Gregory, L., Rimmer, A., Walker, N. M., Yang, T. P., Adlard, J. W., Barwell, J., Berg, J., Brady, A. F., Brewer, C., Brice, G., Chapman, C., Cook, J., Davidson, R., Donaldson, A., Douglas, F., Eccles, D., Evans, D. G., Greenhalgh, L., Henderson, A., Izatt, L., Kumar, A., Lalloo, F., Miedzybrodzka, Z., Morrison, P. J., Paterson, J., Porteous, M., Rogers, M. T., Shanley, S., Walker, L., Gore, M., Houlston, R., Brown, M. A., Caufield, M. J., Deloukas, P., McCarthy, M. I., Todd, J. A., Turnbull, C., Reis-Filho, J. S., Ashworth, A., Antoniou, A. C., Lord, C. J., Donnelly, P. and Rahman, N. (2013). "Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer." Nature **493**(7432): 406-410.

Sadeghi, S., Bain, C. J., Pandeya, N., Webb, P. M., Green, A. C. and Whiteman, D. C. (2008). "Aspirin, nonsteroidal anti-inflammatory drugs, and the risks of cancers of the esophagus." Cancer Epidemiol Biomarkers Prev **17**(5): 1169-1178.

Saito-Ohara, F., Imoto, I., Inoue, J., Hosoi, H., Nakagawara, A., Sugimoto, T. and Inazawa, J. (2003). "PPM1D is a potential target for 17q gain in neuroblastoma." Cancer Research **63**(8): 1876-1883.

Sakakura, C., Hagiwara, A., Miyagawa, K., Nakashima, S., Yoshikawa, T., Kin, S., Nakase, Y., Ito, K., Yamagishi, H., Yazumi, S., Chiba, T. and Ito, Y. (2005). "Frequent downregulation of the runt domain transcription factors RUNX1, RUNX3 and their cofactor CBFB in gastric cancer." Int J Cancer **113**(2): 221-228.

Sammon, A. M. (1992). "A case-control study of diet and social factors in cancer of the esophagus in Transkei." Cancer **69**(4): 860-865.

Satvinder, S. M. and Kang, J.-Y. (2008). Epidemiology and clinical presentation in esophageal cancer. Cancer of the esophagus. Rankin, S., C. Cambridge, Cambridge University Press**:** 1-13.

Saunders, A. E. and Johnson, P. (2010). "Modulation of immune cell signalling by the leukocyte common tyrosine phosphatase, CD45." Cell Signal **22**(3): 339-348.

Savage, S. A., Abnet, C. C., Haque, K., Mark, S. D., Qiao, Y. L., Dong, Z. W., Dawsey, S. M., Taylor, P. R. and Chanock, S. J. (2004). "Polymorphisms in interleukin -2, -6, and -10 are not associated with gastric cardia or esophageal cancer in a high-risk chinese population." Cancer Epidemiol Biomarkers Prev **13**(9): 1547-1549.

Savai, R., Pullamsetti, S. S., Banat, G. A., Weissmann, N., Ghofrani, H. A., Grimminger, F. and Schermuly, R. T. (2010). "Targeting cancer with phosphodiesterase inhibitors." Expert Opin Investig Drugs **19**(1): 117-131.

Schlebusch, C. M., Skoglund, P., Sjodin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G., Soodyall, H. and Jakobsson, M. (2012). "Genomic variation in seven Khoe-San groups reveals adaptation and complex African history." Science **338**(6105): 374-379.

Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., Harris, R. S., Petersen, D. C., Zhao, F., Qi, J., Alkan, C., Kidd, J. M., Sun, Y., Drautz, D. I., Bouffard, P., Muzny, D. M., Reid, J. G., Nazareth, L. V., Wang, Q., Burhans, R., Riemer, C., Wittekindt, N. E., Moorjani, P., Tindall, E. A., Danko, C. G., Teo, W. S., Buboltz, A. M., Zhang, Z., Ma, Q., Oosthuysen, A., Steenkamp, A. W., Oostuisen, H., Venter, P., Gajewski, J., Zhang, Y., Pugh, B. F., Makova, K. D., Nekrutenko, A., Mardis, E. R., Patterson, N., Pringle, T. H., Chiaromonte, F., Mullikin, J. C., Eichler, E. E., Hardison, R. C., Gibbs, R. A., Harkins, T. T. and Hayes, V. M. (2010). "Complete Khoisan and Bantu genomes from southern Africa." Nature **463**(7283): 943-947.

Schwickart, M., Huang, X., Lill, J. R., Liu, J., Ferrando, R., French, D. M., Maecker, H., O'Rourke, K., Bazan, F., Eastham-Anderson, J., Yue, P., Dornan, D., Huang, D. C. and Dixit, V. M. (2010). "Deubiquitinase USP9X stabilizes MCL1 and promotes tumour cell survival." Nature **463**(7277): 103-107.

Sen, G. L., Boxer, L. D., Webster, D. E., Bussat, R. T., Qu, K., Zarnegar, B. J., Johnston, D., Siprashvili, Z. and Khavari, P. A. (2012). "ZNF750 Is a p63 target gene that induces KLF4 to drive terminal epidermal differentiation." Dev Cell **22**(3): 669-677.

Serefoglou, Z., Yapijakis, C., Nkenke, E. and Vairaktaris, E. (2008). "Genetic association of cytokine DNA polymorphisms with head and neck cancer." Oral Oncol **44**(12): 1093-1099.

Shah, T. S., Liu, J. Z., Floyd, J. A., Morris, J. A., Wirth, N., Barrett, J. C. and Anderson, C. A. (2012). "optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants." Bioinformatics **28**(12): 1598-1603.

Shamma, A., Yamamoto, H., Doki, Y., Okami, J., Kondo, M., Fujiwara, Y., Yano, M., Inoue, M., Matsuura, N., Shiozaki, H. and Monden, M. (2000). "Up-

regulation of cyclooxygenase-2 in squamous carcinogenesis of the esophagus." Clin Cancer Res **6**(4): 1229-1238.

Shehadeh, L. A., Yu, K., Wang, L., Guevara, A., Singer, C., Vance, J. and Papapetropoulos, S. (2010). "SRRM2, a potential blood biomarker revealing high alternative splicing in Parkinson's disease." PLoS One **5**(2): e9104.

Siassi, F. and Ghadirian, P. (2005). "Riboflavin deficiency and esophageal cancer: a case control-household study in the Caspian Littoral of Iran." Cancer Detect Prev **29**(5): 464-469.

Siassi, F., Pouransari, Z. and Ghadirian, P. (2000). "Nutrient intake and esophageal cancer in the Caspian littoral of Iran: a case-control study." Cancer Detect Prev **24**(3): 295-303.

Slatkin, M. (2008). "Linkage disequilibrium--understanding the evolutionary past and mapping the medical future." Nat Rev Genet **9**(6): 477-485.

Smith, M., Zhou, M., Whitlock, G., Yang, G., Offer, A., Hui, G., Peto, R., Huang, Z. and Chen, Z. (2008). "Esophageal cancer and body mass index: results from a prospective study of 220,000 men in China and a meta-analysis of published studies." Int J Cancer **122**(7): 1604-1610.

Smith, S. A., Easton, D. F., Evans, D. G. R. and Ponder, B. A. J. (1992). "Allele losses in the region 17q12-21 in familial breast and ovarian cancer involve the wild-type Chromosome." Nat Genet **2**(2): 128-131.

Sobolewski, C., Cerella, C., Dicato, M., Ghibelli, L. and Diederich, M. (2010). "The role of cyclooxygenase-2 in cell proliferation and cell death in human malignancies." Int J Cell Biol **2010**: 215158.

Somdyala, N. I., Bradshaw, D., Gelderblom, W. C. and Parkin, D. M. (2010). "Cancer incidence in a rural population of South Africa, 1998-2002." Int J Cancer **127**(10): 2420-2429.

Somdyala, N. I., Marasas, W. F., Venter, F. S., Vismer, H. F., Gelderblom, W. C. and Swanevelder, S. A. (2003). "Cancer patterns in four districts of the Transkei region--1991-1995." S Afr Med J **93**(2): 144-148.

Song, M. S., Salmena, L., Carracedo, A., Egia, A., Lo-Coco, F., Teruya-Feldstein, J. and Pandolfi, P. P. (2008). "The deubiquitinylation and localization of PTEN are regulated by a HAUSP-PML network." Nature **455**(7214): 813-817.

Song, S., Nones, K., Miller, D., Harliwong, I., Kassahn, K. S., Pinese, M., Pajic, M., Gill, A. J., Johns, A. L., Anderson, M., Holmes, O., Leonard, C., Taylor, D., Wood, S., Xu, Q., Newell, F., Cowley, M. J., Wu, J., Wilson, P., Fink, L., Biankin, A. V., Waddell, N., Grimmond, S. M. and Pearson, J. V. (2012). "qpure: A tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles." PLoS One **7**(9): e45835.

Sorli, S. C., Bunney, T. D., Sugden, P. H., Paterson, H. F. and Katan, M. (2005). "Signaling properties and expression in normal and tumor tissues of two phospholipase C epsilon splice variants." Oncogene **24**(1): 90-100.

Stransky, N., Egloff, A. M., Tward, A. D., Kostic, A. D., Cibulskis, K., Sivachenko, A., Kryukov, G. V., Lawrence, M. S., Sougnez, C., McKenna, A., Shefler, E., Ramos, A. H., Stojanov, P., Carter, S. L., Voet, D., Cortes, M. L., Auclair, D., Berger, M. F., Saksena, G., Guiducci, C., Onofrio, R. C., Parkin, M., Romkes, M., Weissfeld, J. L., Seethala, R. R., Wang, L., Rangel-Escareno, C., Fernandez-Lopez, J. C., Hidalgo-Miranda, A., Melendez-Zajgla, J., Winckler, W., Ardlie, K., Gabriel, S. B., Meyerson, M., Lander, E. S., Getz, G., Golub, T. R., Garraway, L. A. and Grandis, J. R. (2011). "The mutational landscape of head and neck squamous cell carcinoma." Science **333**(6046): 1157-1160.

Strasak, A. M., Pfeiffer, R. M., Brant, L. J., Rapp, K., Hilbe, W., Oberaigner, W., Lang, S., Borena, W., Concin, H., Diem, G., Ruttmann, E., Glodny, B., Pfeiffer, K. P. and Ulmer, H. (2009). "Time-dependent association of total serum cholesterol and cancer incidence in a cohort of 172,210 men and women: a prospective 19-year follow-up study." Ann Oncol **20**(6): 1113-1120.

Stratton, M. R. (2011). "Exploring the genomes of cancer cells: progress and promise." Science **331**(6024): 1553-1558.

Streppel, M. M., Lata, S., Delabastide, M., Montgomery, E. A., Wang, J. S., Canto, M. I., Macgregor-Das, A. M., Pai, S., Morsink, F. H., Offerhaus, G. J., Antoniou, E., Maitra, A. and McCombie, W. R. (2013). "Next-generation sequencing of endoscopic biopsies identifies ARID1A as a tumor-suppressor gene in Barrett's esophagus." Oncogene **[Epub ahead of print]**.

Su, Z., Gay, L. J., Strange, A., Palles, C., Band, G., Whiteman, D. C., Lescai, F., Langford, C., Nanji, M., Edkins, S., van der Winkel, A., Levine, D., Sasieni, P., Bellenguez, C., Howarth, K., Freeman, C., Trudgill, N., Tucker, A. T., Pirinen, M., Peppelenbosch, M. P., van der Laan, L. J. W., Kuipers, E. J., Drenth, J. P. H., Peters, W. H., Reynolds, J. V., Kelleher, D. P., McManus, R., Grabsch, H., Prenen, H., Bisschops, R., Krishnadath, K., Siersema, P. D., van Baal, J. W. P. M., Middleton, M.,

Petty, R., Gillies, R., Burch, N., Bhandari, P., Paterson, S., Edwards, C., Penman, I., Vaidya, K., Ang, Y., Murray, I., Patel, P., Ye, W. M., Mullins, P., Wu, A. H., Bird, N. C., Dallal, H., Shaheen, N. J., Murray, L. J., Koss, K., Bernstein, L., Romero, Y., Hardie, L. J., Zhang, R., Winter, H., Corley, D. A., Panter, S., Risch, H. A., Reid, B. J., Sargeant, I., Gammon, M. D., Smart, H., Dhar, A., McMurtry, H., Ali, H., Liu, G., Casson, A. G., Chow, W. H., Rutter, M., Tawil, A., Morris, D., Nwokolo, C., Isaacs, P., Rodgers, C., Ragunath, K., MacDonald, C., Haigh, C., Monk, D., Davies, G., Wajed, S., Johnston, D., Gibbons, M., Cullen, S., Church, N., Langley, R., Griffin, M., Alderson, D., Deloukas, P., Hunt, S. E., Gray, E., Dronov, S., Potter, S. C., Tashakkori-Ghanbaria, A., Anderson, M., Brooks, C., Blackwell, J. M., Bramon, E., Brown, M. A., Casas, J. P., Corvin, A., Duncanson, A., Markus, H. S., Mathew, C. G., Palmer, C. N. A., Plomin, R., Rautanen, A., Sawcer, S. J., Trembath, R. C., Viswanathan, A. C., Wood, N., Trynka, G., Wijmenga, C., Cazier, J. B., Atherfold, P., Nicholson, A. M., Gellatly, N. L., Glancy, D., Cooper, S. C., Cunningham, D., Lind, T., Hapeshi, J., Ferry, D., Rathbone, B., Brown, J., Love, S., Attwood, S., MacGregor, S., Watson, P., Sanders, S., Ek, W., Harrison, R. F., Moayyedi, P., de Caestecker, J., Barr, H., Stupka, E., Vaughan, T. L., Peltonen, L., Spencer, C. C. A., Tomlinson, I., Donnelly, P., Jankowski, J. A. Z., Esophageal Adenocarcinoma Genetics Consortium and 2., W. T. C. C. C. (2012). "Common variants at the MHC locus and at chromosome 16q24.1 predispose to Barrett's esophagus." Nat Genet **44**(10): 1131-1136.

Subramanian, V. S., Rapp, L., Marchant, J. S. and Said, H. M. (2011). "Role of cysteine residues in cell surface expression of the human riboflavin transporter-2 (hRFT2) in intestinal epithelial cells." Am J Physiol Gastrointest Liver Physiol **301**(1): G100-G109.

Sun, T., Gao, Y., Tan, W., Ma, S., Shi, Y., Yao, J., Guo, Y., Yang, M., Zhang, X., Zhang, Q., Zeng, C. and Lin, D. (2007). "A six-nucleotide insertion-deletion polymorphism in the CASP8 promoter is associated with susceptibility to multiple cancers." Nat Genet **39**(5): 605-613.

Sun, T., Miao, X., Zhang, X., Tan, W., Xiong, P. and Lin, D. (2004). "Polymorphisms of death pathway genes FAS and FASL in esophageal squamous-cell carcinoma." J Natl Cancer Inst **96**(13): 1030-1036.

Suzuki, H., Zhou, X., Yin, J., Lei, J., Jiang, H. Y., Suzuki, Y., Chan, T., Hannon, G. J., Mergner, W. J., Abraham, J. M. and Meltzer, S. J. (1995). "Intragenic mutations of CDKN2B and CDKN2A in primary human esophageal cancers." Hum Mol Genet **4**(10): 1883-1887.

Sydenham, E. W., Thiel, P. G., Marasas, W. F. O., Shephard, G. S., Vanschalkwyk, D. J. and Koch, K. R. (1990). "Natural occurrence of

some fusarium mycotoxins in corn from low and high esophageal cancer prevalence areas of the Transkei, southern Africa." J Agric Food Chem **38**(10): 1900-1903.

Szczeklik, W., Sanak, M. and Szczeklik, A. (2004). "Functional effects and gender association of COX-2 gene polymorphism G-765C in bronchial asthma." J Allergy Clin Immunol **114**(2): 248-253.

Taguchi-Atarashi, N., Hamasaki, M., Matsunaga, K., Omori, H., Ktistakis, N. T., Yoshimori, T. and Noda, T. (2010). "Modulation of local PtdIns3P levels by the PI phosphatase MTMR3 regulates constitutive autophagy." Traffic **11**(4): 468-478.

Tan, D. S. P., Lambros, M. B. K., Rayter, S., Natrajan, R., Vatcheva, R., Gao, Q., Marchio, C., Geyer, F. C., Savage, K., Parry, S., Fenwick, K., Tamber, N., Mackay, A., Dexter, T., Jameson, C., McCluggage, W. G., Williams, A., Graham, A., Faratian, D., El-Bahrawy, M., Paige, A. J., Gabra, H., Gore, M. E., Zvelebil, M., Lord, C. J., Kaye, S. B., Ashworth, A. and Reis, J. S. (2009). "PPM1D Is a Potential Therapeutic Target in Ovarian Clear Cell Carcinomas." Clin Cancer Res **15**(7): 2269-2280.

Tanaka, F., Yamamoto, K., Suzuki, S., Inoue, H., Tsurumaru, M., Kajiyama, Y., Kato, H., Igaki, H., Furuta, K., Fujita, H., Tanaka, T., Tanaka, Y., Kawashima, Y., Natsugoe, S., Setoyama, T., Tokudome, S., Mimori, K., Haraguchi, N., Ishii, H. and Mori, M. (2010). "Strong interaction between the effects of alcohol consumption and smoking on oesophageal squamous cell carcinoma among individuals with ADH1B and/or ALDH2 risk alleles." Gut **59**(11): 1457-1464.

Tao, Y. P., Wang, W. L., Li, S. Y., Zhang, J., Shi, Q. Z., Zhao, F. and Zhao, B. S. (2012). "Associations between polymorphisms in IL-12A, IL-12B, IL-12R beta 1, IL-27 gene and serum levels of IL-12p40, IL-27p28 with esophageal cancer." J Cancer Res Clin Oncol **138**(11): 1891-1900.

Teo, Y. Y., Small, K. S. and Kwiatkowski, D. P. (2010). "Methodological challenges of genome-wide association analysis in Africa." Nat Rev Genet **11**(2): 149-160.

Terry, P., Lagergren, J., Wolk, A. and Nyren, O. (2001). "Drinking hot beverages is not associated with risk of oesophageal cancers in a Western population." Br J Cancer **84**(1): 120-121.

Thye, T., Owusu-Dabo, E., Vannberg, F. O., van Crevel, R., Curtis, J., Sahiratmadja, E., Balabanova, Y., Ehmen, C., Muntau, B., Ruge, G., Sievertsen, J., Gyapong, J., Nikolayevskyy, V., Hill, P. C., Sirugo, G., Drobniewski, F., van de Vosse, E., Newport, M., Alisjahbana, B.,

Nejentsev, S., Ottenhoff, T. H. M., Hill, A. V. S., Horstmann, R. D. and Meyer, C. G. (2012). "Common variants at 11p13 are associated with susceptibility to tuberculosis." Nat Genet **44**(3): 257-259.

Timmann, C., Thye, T., Vens, M., Evans, J., May, J., Ehmen, C., Sievertsen, J., Muntau, B., Ruge, G., Loag, W., Ansong, D., Antwi, S., Asafo-Adjei, E., Nguah, S. B., Kwakye, K. O., Akoto, A. O., Sylverken, J., Brendel, M., Schuldt, K., Loley, C., Franke, A., Meyer, C. G., Agbenyega, T., Ziegler, A. and Horstmann, R. D. (2012). "Genome-wide association study indicates two novel resistance loci for severe malaria." Nature **489**(7416): 443-446.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J. M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L. and Williams, S. M. (2009). "The genetic structure and history of Africans and African Americans." Science **324**(5930): 1035-1044.

Tishkoff, S. A. and Williams, S. M. (2002). "Genetic analysis of African populations: human evolution and complex disease." Nat Rev Genet **3**(8): 611-621.

Toh, Y., Oki, E., Ohgaki, K., Sakamoto, Y., Ito, S., Egashira, A., Saeki, H., Kakeji, Y., Morita, M., Sakaguchi, Y., Okamura, T. and Maehara, Y. (2010). "Alcohol drinking, cigarette smoking, and the development of squamous cell carcinoma of the esophagus: molecular mechanisms of carcinogenesis." Int J Clin Oncol **15**(2): 135-144.

Tonomoto, Y., Tachibana, M., Dhar, D. K., Onoda, T., Hata, K., Ohnuma, H., Tanaka, T. and Nagasue, N. (2007). "Differential expression of RUNX genes in human esophageal squamous cell carcinoma: downregulation of RUNX3 worsens patient prognosis." Oncology **73**(5-6): 346-356.

Trynka, G., Hunt, K. A., Bockett, N. A., Romanos, J., Mistry, V., Szperl, A., Bakker, S. F., Bardella, M. T., Bhaw-Rosun, L., Castillejo, G., de la Concha, E. G., de Almeida, R. C., Dias, K. R., van Diemen, C. C., Dubois, P. C., Duerr, R. H., Edkins, S., Franke, L., Fransen, K., Gutierrez, J., Heap, G. A., Hrdlickova, B., Hunt, S., Izurieta, L. P., Izzo, V., Joosten, L. A., Langford, C., Mazzilli, M. C., Mein, C. A., Midah, V., Mitrovic, M., Mora, B., Morelli, M., Nutland, S., Nunez, C., Onengut-Gumuscu, S., Pearce, K., Platteel, M., Polanco, I., Potter, S., Ribes-Koninckx, C., Ricano-Ponce, I., Rich, S. S., Rybak, A., Santiago, J. L., Senapati, S., Sood, A., Szajewska, H., Troncone, R., Varade, J., Wallace, C., Wolters, V. M., Zhernakova, A., Thelma, B. K., Cukrowska,

B., Urcelay, E., Bilbao, J. R., Mearin, M. L., Barisani, D., Barrett, J. C., Plagnol, V., Deloukas, P., Wijmenga, C. and van Heel, D. A. (2011). "Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease." Nat Genet **43**(12): 1193-1201.

Tsoi, L. C., Spain, S. L., Knight, J., Ellinghaus, E., Stuart, P. E., Capon, F., Ding, J., Li, Y., Tejasvi, T., Gudjonsson, J. E., Kang, H. M., Allen, M. H., McManus, R., Novelli, G., Samuelsson, L., Schalkwijk, J., Stahle, M., Burden, A. D., Smith, C. H., Cork, M. J., Estivill, X., Bowcock, A. M., Krueger, G. G., Weger, W., Worthington, J., Tazi-Ahnini, R., Nestle, F. O., Hayday, A., Hoffmann, P., Winkelmann, J., Wijmenga, C., Langford, C., Edkins, S., Andrews, R., Blackburn, H., Strange, A., Band, G., Pearson, R. D., Vukcevic, D., Spencer, C. C., Deloukas, P., Mrowietz, U., Schreiber, S., Weidinger, S., Koks, S., Kingo, K., Esko, T., Metspalu, A., Lim, H. W., Voorhees, J. J., Weichenthal, M., Wichmann, H. E., Chandran, V., Rosen, C. F., Rahman, P., Gladman, D. D., Griffiths, C. E., Reis, A., Kere, J., Duffin, K. C., Helms, C., Goldgar, D., Paschall, J., Malloy, M. J., Pullinger, C. R., Kane, J. P., Gardner, J., Perlmutter, A., Miner, A., Feng, B. J., Hiremagalore, R., Ike, R. W., Christophers, E., Henseler, T., Ruether, A., Schrodi, S. J., Prahalad, S., Guthery, S. L., Fischer, J., Liao, W., Kwok, P., Menter, A., Lathrop, G. M., Wise, C., Begovich, A. B., Onoufriadis, A., Weale, M. E., Hofer, A., Salmhofer, W., Wolf, P., Kainu, K., Saarialho-Kere, U., Suomela, S., Badorf, P., Huffmeier, U., Kurrat, W., Kuster, W., Lascorz, J., Mossner, R., Schurmeier-Horst, F., Stander, M., Traupe, H., Bergboer, J. G., Heijer, M., van de Kerkhof, P. C., Zeeuwen, P. L., Barnes, L., Campbell, L. E., Cusack, C., Coleman, C., Conroy, J., Ennis, S., Fitzgerald, O., Gallagher, P., Irvine, A. D., Kirby, B., Markham, T., McLean, W. H., McPartlin, J., Rogers, S. F., Ryan, A. W., Zawirska, A., Giardina, E., Lepre, T., Perricone, C., Martin-Ezquerra, G., Pujol, R. M., Riveira-Munoz, E., Inerot, A., Naluai, A. T., Mallbris, L., Wolk, K., Leman, J., Barton, A., Warren, R. B., Young, H. S., Ricano-Ponce, I., Trynka, G., Pellett, F. J., Henschel, A., Aurand, M., Bebo, B., Gieger, C., Illig, T., Moebus, S., Jockel, K. H., Erbel, R., Donnelly, P., Peltonen, L., Blackwell, J. M., Bramon, E., Brown, M. A., Casas, J. P., Corvin, A., Craddock, N., Duncanson, A., Jankowski, J., Markus, H. S., Mathew, C. G., McCarthy, M. I., Palmer, C. N., Plomin, R., Rautanen, A., Sawcer, S. J., Samani, N., Viswanathan, A. C., Wood, N. W., Bellenguez, C., Freeman, C., Hellenthal, G., Giannoulatou, E., Pirinen, M., Su, Z., Hunt, S. E., Gwilliam, R., Bumpstead, S. J., Dronov, S., Gillman, M., Gray, E., Hammond, N., Jayakumar, A., McCann, O. T., Liddle, J., Perez, M. L., Potter, S. C., Ravindrarajah, R., Ricketts, M., Waller, M., Weston, P., Widaa, S., Whittaker, P., Nair, R. P., Franke, A., Barker, J. N., Abecasis, G. R., Elder, J. T. and Trembath, R. C. (2012). "Identification of 15 new

psoriasis susceptibility loci highlights the role of innate immunity." Nat Genet **44**(12): 1341-1348.

Turley, R. S., Finger, E. C., Hempel, N., How, T., Fields, T. A. and Blobe, G. C. (2007). "The type III transforming growth factor-beta receptor as a novel tumor suppressor gene in prostate cancer." Cancer Res **67**(3): 1090-1098.

U.S. Department of Health and Human Services (2010). How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General. Atlanta, GA, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. Available: http://www.surgeongeneral.gov/library/reports/tobaccosmoke/index.html.

Umar, M., Upadhyay, R., Kumar, S., Ghoshal, U. C. and Mittal, B. (2011). "CASP8 -652 6N del and CASP8 IVS12-19G>A gene polymorphisms and susceptibility/prognosis of ESCC: a case control study in northern Indian population." J Surg Oncol **103**(7): 716-723.

Umar, S. B. and Fleischer, D. E. (2008). "Esophageal cancer: epidemiology, pathogenesis and prevention." Nat Clin Pract Gastroenterol Hepatol **5**(9): 517-526.

Upadhyay, R., Jain, M., Kumar, S., Ghoshal, U. C. and Mittal, B. (2008). "Association of interleukin-6 (-174G>C) promoter polymorphism with risk of squamous cell esophageal cancer and tumor location: an exploratory study." Clin Immunol **128**(2): 199-204.

Upadhyay, R., Jain, M., Kumar, S., Ghoshal, U. C. and Mittal, B. (2009). "Functional polymorphisms of cyclooxygenase-2 (COX-2) gene and risk for esophageal squmaous cell carcinoma." Mutat Res **663**(1-2): 52-59.

Urano, T., Shiraki, M., Yagi, H., Ito, M., Sasaki, N., Sato, M., Ouchi, Y. and Inoue, S. (2012). "GPR98/Gpr98 gene is involved in the regulation of human and mouse bone mineral density." J Clin Endocrinol Metab **97**(4): E565-E574.

Varghese, J. S. and Easton, D. F. (2010). "Genome-wide association studies in common cancers--what have we learnt?" Curr Opin Genet Dev **20**(3): 201-209.

Vivanco, I. and Sawyers, C. L. (2002). "The phosphatidylinositol 3-Kinase AKT pathway in human cancer." Nat Rev Cancer **2**(7): 489-501.

Vogelsang, M., Wang, Y., Veber, N., Mwapagha, L. M. and Parker, M. I. (2012). "The cumulative effects of polymorphisms in the DNA mismatch repair genes and tobacco smoking in oesophageal cancer risk." PLoS One **7**(5): e36962.

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., Burtt, N. P., Fuchsberger, C., Li, Y., Erdmann, J., Frayling, T. M., Heid, I. M., Jackson, A. U., Johnson, T., Kilpelainen, T. O., Lindgren, C. M., Morris, A. P., Prokopenko, I., Randall, J. C., Saxena, R., Soranzo, N., Speliotes, E. K., Teslovich, T. M., Wheeler, E., Maguire, J., Parkin, M., Potter, S., Rayner, N. W., Robertson, N., Stirrups, K., Winckler, W., Sanna, S., Mulas, A., Nagaraja, R., Cucca, F., Barroso, I., Deloukas, P., Loos, R. J., Kathiresan, S., Munroe, P. B., Newton-Cheh, C., Pfeufer, A., Samani, N. J., Schunkert, H., Hirschhorn, J. N., Altshuler, D., McCarthy, M. I., Abecasis, G. R. and Boehnke, M. (2012). "The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits." PLoS Genet **8**(8): e1002793.

Vousden, K. H. and Lu, X. (2002). "Live or let die: the cell's response to p53." Nat Rev Cancer **2**(8): 594-604.

Wabinga, H. R., Parkin, D. M., Wabwire-Mangen, F. and Nambooze, S. (2000). "Trends in cancer incidence in Kyadondo County, Uganda, 1960-1997." Br J Cancer **82**(9): 1585-1592.

Wang, F. L., Wang, Y., Wong, W. K., Liu, Y., Addivinola, F. J., Liang, P., Chen, L. B., Kantoff, P. W. and Pardee, A. B. (1996). "Two differentially expressed genes in normal human prostate tissue and in carcinoma." Cancer Res **56**(16): 3634-3637.

Wang, K., Li, M. Y. and Hakonarson, H. (2010.b). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." Nucleic Acids Res **38**(16): e164.

Wang, L., Wang, X., Jie, P., Lu, H., Zhang, S., Lin, X., Lam, E. K., Cui, Y., Yu, J. and Jin, H. (2011). "Klotho is silenced through promoter hypermethylation in gastric cancer." Am J Cancer Res **1**(1): 111-119.

Wang, L. D., Zhou, F. Y., Li, X. M., Sun, L. D., Song, X., Jin, Y., Li, J. M., Kong, G. Q., Qi, H., Cui, J., Zhang, L. Q., Yang, J. Z., Li, J. L., Li, X. C., Ren, J. L., Liu, Z. C., Gao, W. J., Yuan, L., Wei, W., Zhang, Y. R., Wang, W. P., Sheyhidin, I., Li, F., Chen, B. P., Ren, S. W., Liu, B., Li, D., Ku, J. W., Fan, Z. M., Zhou, S. L., Guo, Z. G., Zhao, X. K., Liu, N., Ai, Y. H., Shen, F. F., Cui, W. Y., Song, S., Guo, T., Huang, J., Yuan, C., Wu, Y., Yue, W. B., Feng, C. W., Li, H. L., Wang, Y., Tian, J. Y., Lu, Y., Yuan, Y., Zhu, W. L., Liu, M., Fu, W. J., Yang, X., Wang, H. J., Han, S. L., Chen, J., Han,

M., Wang, H. Y., Zhang, P., Dong, J. C., Xing, G. L., Wang, R., Guo, M., Chang, Z. W., Liu, H. L., Guo, L., Yuan, Z. Q., Liu, H., Lu, Q., Yang, L. Q., Zhu, F. G., Yang, X. F., Feng, X. S., Wang, Z., Li, Y., Gao, S. G., Qige, Q., Bai, L. T., Yang, W. J., Lei, G. Y., Shen, Z. Y., Chen, L. Q., Li, E. M., Xu, L. Y., Wu, Z. Y., Cao, W. K., Wang, J. P., Bao, Z. Q., Chen, J. L., Ding, G. C., Zhuang, X., Zhou, Y. F., Zheng, H. F., Zhang, Z., Zuo, X. B., Dong, Z. M., Fan, D. M., He, X., Wang, J., Zhou, Q., Zhang, Q. X., Jiao, X. Y., Lian, S. Y., Ji, A. F., Lu, X. M., Wang, J. S., Chang, F. B., Lu, C. D., Chen, Z. G., Miao, J. J., Fan, Z. L., Lin, R. B., Liu, T. J., Wei, J. C., Kong, Q. P., Lan, Y., Fan, Y. J., Gao, F. S., Wang, T. Y., Xie, D., Chen, S. Q., Yang, W. C., Hong, J. Y., Wang, L., Qiu, S. L., Cai, Z. M. and Zhang, X. J. (2010.a). "Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies susceptibility loci at PLCE1 and C20orf54." Nat Genet **42**(9): 759-763.

Wang, M., Zhang, Z., Tian, Y. and Shao, J. (2009). "A six-nucleotide insertion-deletion polymorphism in the CASP8 promoter associated with risk and progression of bladder cancer." Clin Cancer Res **15**(7): 2567-2572.

Wang, X., Zbou, C., Qiu, G., Fan, J., Tang, H. and Peng, Z. (2008). "Screening of new tumor suppressor genes in sporadic colorectal cancer patients." Hepatogastroenterology **55**(88): 2039-2044.

Wang, X., Zhou, C., Qiu, G., Yang, Y., Yan, D., Xing, T., Fan, J., Tang, H. and Peng, Z. (2012). "Phospholipase C epsilon plays a suppressive role in incidence of colorectal cancer." Med Oncol **29**(2): 1051-1058.

Ward, L. D. and Kellis, M. (2012). "HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants." Nucleic Acids Res **40**(Database issue): D930-934.

Warr, M. R. and Shore, G. C. (2008). "Unique biology of Mcl-1: therapeutic opportunities in cancer." Curr Mol Med **8**(2): 138-147.

Waters, K. M., Stram, D. O., Hassanein, M. T., Le Marchand, L., Wilkens, L. R., Maskarinec, G., Monroe, K. R., Kolonel, L. N., Altshuler, D., Henderson, B. E. and Haiman, C. A. (2010). "Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups." PLoS Genet **6**(8): e1001078.

Weir, B. A., Woo, M. S., Getz, G., Perner, S., Ding, L., Beroukhim, R., Lin, W. M., Province, M. A., Kraja, A., Johnson, L. A., Shah, K., Sato, M., Thomas, R. K., Barletta, J. A., Borecki, I. B., Broderick, S., Chang, A. C., Chiang, D. Y., Chirieac, L. R., Cho, J., Fujii, Y., Gazdar, A. F., Giordano, T., Greulich, H., Hanna, M., Johnson, B. E., Kris, M. G., Lash, A., Lin, L.,

Lindeman, N., Mardis, E. R., McPherson, J. D., Minna, J. D., Morgan, M. B., Nadel, M., Orringer, M. B., Osborne, J. R., Ozenberger, B., Ramos, A. H., Robinson, J., Roth, J. A., Rusch, V., Sasaki, H., Shepherd, F., Sougnez, C., Spitz, M. R., Tsao, M. S., Twomey, D., Verhaak, R. G., Weinstock, G. M., Wheeler, D. A., Winckler, W., Yoshizawa, A., Yu, S., Zakowski, M. F., Zhang, Q., Beer, D. G., Wistuba, II, Watson, M. A., Garraway, L. A., Ladanyi, M., Travis, W. D., Pao, W., Rubin, M. A., Gabriel, S. B., Gibbs, R. A., Varmus, H. E., Wilson, R. K., Lander, E. S. and Meyerson, M. (2007). "Characterizing the cancer genome in lung adenocarcinoma." Nature **450**(7171): 893-898.

Wen, D., Wang, S., Zhang, L., Wei, L., Zhou, W. and Peng, Q. (2009). "Early onset, multiple primary malignancies, and poor prognosis are indicative of an inherited predisposition to esophageal squamous cell carcinoma for the familial as opposed to the sporadic cases--an update on over 14-year survival." Eur J Med Genet **52**(6): 381-385.

Wen, D., Wang, S., Zhang, L., Zhang, J., Wei, L. and Zhao, X. (2006). "Differences of onset age and survival rates in esophageal squamous cell carcinoma cases with and without family history of upper gastrointestinal cancer from a high-incidence area in North China." Fam Cancer **5**(4): 343-352.

Westbrook, T. F., Martin, E. S., Schlabach, M. R., Leng, Y., Liang, A. C., Feng, B., Zhao, J. J., Roberts, T. M., Mandel, G., Hannon, G. J., Depinho, R. A., Chin, L. and Elledge, S. J. (2005). "A genetic screen for candidate tumor suppressors identifies REST." Cell **121**(6): 837-848.

Weston, M. D., Luijendijk, M. W., Humphrey, K. D., Moller, C. and Kimberling, W. J. (2004). "Mutations in the VLGR1 gene implicate G-protein signaling in the pathogenesis of Usher syndrome type II." Am J Hum Genet **74**(2): 357-366.

White, E., Karp, C., Strohecker, A. M., Guo, Y. and Mathew, R. (2010). "Role of autophagy in suppression of inflammation and cancer." Curr Opin Cell Biol **22**(2): 212-217.

Whiteman, D. C., Sadeghi, S., Pandeya, N., Smithers, B. M., Gotley, D. C., Bain, C. J., Webb, P. M. and Green, A. C. (2008). "Combined effects of obesity, acid reflux and smoking on the risk of adenocarcinomas of the oesophagus." Gut **57**(2): 173-180.

Winkler, C. A., Nelson, G. W. and Smith, M. W. (2010). "Admixture mapping comes of age." Annu Rev Genomics Hum Genet **11**: 65-89.

World Health Organization (2008). The global burden of disease: 2004 update. Geneva, Switzerland, World Health Organization. http://www.who.int/healthinfo/global_burden_disease/GBD_report_2004update_full.pdf.

Wu, C., Hu, Z., He, Z., Jia, W., Wang, F., Zhou, Y., Liu, Z., Zhan, Q., Liu, Y., Yu, D., Zhai, K., Chang, J., Qiao, Y., Jin, G., Shen, Y., Guo, C., Fu, J., Miao, X., Tan, W., Shen, H., Ke, Y., Zeng, Y., Wu, T. and Lin, D. (2011.c). "Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations." Nat Genet **43**(7): 679-684.

Wu, C., Kraft, P., Zhai, K., Chang, J., Wang, Z., Li, Y., Hu, Z., He, Z., Jia, W., Abnet, C. C., Liang, L., Hu, N., Miao, X., Zhou, Y., Liu, Z., Zhan, Q., Liu, Y., Qiao, Y., Jin, G., Guo, C., Lu, C., Yang, H., Fu, J., Yu, D., Freedman, N. D., Ding, T., Tan, W., Goldstein, A. M., Wu, T., Shen, H., Ke, Y., Zeng, Y., Chanock, S. J., Taylor, P. R. and Lin, D. (2012.a). "Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions." Nat Genet **44**(10): 1090-1097.

Wu, I. C., Zhao, Y., Zhai, R., Liu, C. Y., Chen, F., Ter-Minassian, M., Asomaning, K., Su, L., Heist, R. S., Kulke, M. H., Liu, G. and Christiani, D. C. (2011.a). "Interactions between genetic polymorphisms in the apoptotic pathway and environmental factors on esophageal adenocarcinoma risk." Carcinogenesis **32**(4): 502-506.

Wu, J., Metz, C., Xu, X., Abe, R., Gibson, A. W., Edberg, J. C., Cooke, J., Xie, F., Cooper, G. S. and Kimberly, R. P. (2003). "A novel polymorphic CAAT/enhancer-binding protein beta element in the FasL gene promoter alters Fas ligand expression: a candidate background gene in African American systemic lupus erythematosus patients." J Immunol **170**(1): 132-138.

Wu, M., Zhang, Z. F., Kampman, E., Zhou, J. Y., Han, R. Q., Yang, J., Zhang, X. F., Gu, X. P., Liu, A. M., van't Veer, P., Kok, F. J. and Zhao, J. K. (2011.b). "Does family history of cancer modify the effects of lifestyle risk factors on esophageal cancer? A population-based case-control study in China." Int J Cancer **128**(9): 2147-2157.

Wu, Y., Wang, X., Wu, F., Huang, R., Xue, F., Liang, G., Tao, M., Cai, P. and Huang, Y. (2012.b). "Transcriptome profiling of the cancer, adjacent non-tumor and distant normal tissues from a colorectal cancer patient by deep sequencing." PLoS One **7**(8): e41001.

Wu, Y. R., Chen, C. M., Chen, Y. C., Chao, C. Y., Ro, L. S., Fung, H. C., Hsiao, Y. C., Hu, F. J. and Lee-Chen, G. J. (2010). "Ubiquitin specific proteases USP24 and USP40 and ubiquitin thiolesterase UCHL1 polymorphisms have synergic effect on the risk of Parkinson's disease among Taiwanese." Clin Chim Acta **411**(13-14): 955-958.

Xiong, D. H., Liu, X. G., Guo, Y. F., Tan, L. J., Wang, L., Sha, B. Y., Tang, Z. H., Pan, F., Yang, T. L., Chen, X. D., Lei, S. F., Yerges, L. M., Zhu, X. Z., Wheeler, V. W., Patrick, A. L., Bunker, C. H., Guo, Y., Yan, H., Pei, Y. F., Zhang, Y. P., Levy, S., Papasian, C. J., Xiao, P., Lundberg, Y. W., Recker, R. R., Liu, Y. Z., Liu, Y. J., Zmuda, J. M. and Deng, H. W. (2009). "Genome-wide association and follow-up replication studies identified ADAMTS18 and TGFBR3 as bone mass candidate genes in different ethnic groups." Am J Hum Genet **84**(3): 388-398.

Yamaguchi, H., Wyckoff, J. and Condeelis, J. (2005). "Cell migration in tumors." Curr Opin Cell Biol **17**(5): 559-564.

Yin, M., Yan, J., Wei, S. and Wei, Q. (2010). "CASP8 polymorphisms contribute to cancer susceptibility: evidence from a meta-analysis of 23 publications with 55 individual studies." Carcinogenesis **31**(5): 850-857.

Yokoyama, A., Kato, H., Yokoyama, T., Tsujinaka, T., Muto, M., Omori, T., Haneda, T., Kumagai, Y., Igaki, H., Yokoyama, M., Watanabe, H., Fukuda, H. and Yoshimizu, H. (2002). "Genetic polymorphisms of alcohol and aldehyde dehydrogenases and glutathione S-transferase M1 and drinking, smoking, and diet in Japanese men with esophageal squamous cell carcinoma." Carcinogenesis **23**(11): 1851-1859.

Yoshida, A., Hsu, L. C. and Yasunami, M. (1991). "Genetics of human alcohol-metabolizing enzymes." Prog Nucleic Acid Res Mol Biol **40**: 255-287.

Yoshida, A., Huang, I. Y. and Ikawa, M. (1984). "Molecular abnormality of an inactive aldehyde dehydrogenase variant commonly found in Orientals." Proc Natl Acad Sci U S A **81**(1): 258-261.

Yuan, Z., Yuan, H., Ma, H., Chu, M., Wang, Y., Hu, Z., Shen, H. and Chen, N. (2013). "Genetic variants at 10q23 are associated with risk of head and neck cancer in a Chinese population." Oral Oncol **49**(4): 332-335.

Zagari, R. M., Fuccio, L., Wallander, M. A., Johansson, S., Fiocca, R., Casanova, S., Farahmand, B. Y., Winchester, C. C., Roda, E. and Bazzoli, F. (2008). "Gastro-oesophageal reflux symptoms, oesophagitis and Barrett's oesophagus in the general population: the Loiano-Monghidoro study." Gut **57**(10): 1354-1359.

Zhai, R., Zhao, Y., Liu, G., Ter-Minassian, M., Wu, I. C., Wang, Z., Su, L., Asomaning, K., Chen, F., Kulke, M. H., Lin, X., Heist, R. S., Wain, J. C. and Christiani, D. C. (2012). "Interactions between environmental factors and polymorphisms in angiogenesis pathway genes in esophageal adenocarcinoma risk: a case-only study." Cancer **118**(3): 804-811.

Zhang, F., Yang, Y., Guo, C. and Wang, Y. (2012). "CASP8 -652 6N del polymorphism and cancer risk: a meta-analysis of 30 case-control studies in 50,112 subjects." Mutagenesis **27**(5): 559-566.

Zhang, X., Miao, X., Tan, W., Ning, B., Liu, Z., Hong, Y., Song, W., Guo, Y., Shen, Y., Qiang, B., Kadlubar, F. F. and Lin, D. (2005). "Identification of functional genetic variants in cyclooxygenase-2 and their association with risk of esophageal cancer." Gastroenterology **129**(2): 565-576.

Zhang, Z., Qiu, L., Wang, M., Tong, N. and Li, J. (2009.b). "The FAS ligand promoter polymorphism, rs763110 (-844C>T), contributes to cancer susceptibility: evidence from 19 case-control studies." Eur J Hum Genet **17**(10): 1294-1303.

Zhang, Z., Xue, H., Gong, W., Wang, M., Yuan, L. and Han, S. (2009.a). "FAS promoter polymorphisms and cancer risk: a meta-analysis based on 34 case-control studies." Carcinogenesis **30**(3): 487-493.

Zhong, Y., Huang, Y., Zhang, T., Ma, C., Zhang, S., Fan, W., Chen, H., Qian, J. and Lu, D. (2010). "Effects of O6-methylguanine-DNA methyltransferase (MGMT) polymorphisms on cancer: a meta-analysis." Mutagenesis **25**(1): 83-95.

Zhu, Y., McAvoy, S., Kuhn, R. and Smith, D. I. (2006). "RORA, a large common fragile site gene, is involved in cellular stress response." Oncogene **25**(20): 2901-2908.

Zhuo, X., Zhang, Y., Wang, Y., Zhuo, W., Zhu, Y. and Zhang, X. (2008). "Helicobacter pylori infection and oesophageal cancer risk: association studies via evidence-based meta-analyses." Clin Oncol (R Coll Radiol) **20**(10): 757-762.

Zimmermann, K. C., Sarbia, M., Weber, A. A., Borchard, F., Gabbert, H. E. and Schror, K. (1999). "Cyclooxygenase-2 expression in human esophageal carcinoma." Cancer Res **59**(1): 198-204.

Appendix

Table A.1: Primers and PCR conditions for amplification of *PLCE1* exons

Exon	Primer	Sequence	Annealing temp (°C)	Extension time	Product length (bp)
1	Fwd	GGGAGCGGACTGTGAACG	63	30 sec	444
	Rev	GAGCGCGAGGACACTTTTC			
2	Fwd	TTTTGGCTGGAAGCAGAAGT	57	2 min	1922
	Rev	GCTATCAATTGGAGTATCTGTTTTCA			
1*	Fwd	CCTCCCTCGATTCTGGGTAT	58	30 sec	472
	Rev	AACAGTCCCCAGGATTCCAT			
3	Fwd	TTTGCACTTGGAGCATCTGA	55	30 sec	543
	Rev	TTCTCTTAAAGAAGCGACTTTTACTT			
4	Fwd	CAGAACTCTTCACTAAGCAGAGGA	60	30 sec	568
	Rev	CATGGAAGTGGCAAGACAGA			
5	Fwd	CCCAGCCAGGACCTACAG	60	30 sec	437
	Rev	GCCTTTGGTGCTGAAAGAAG			
6	Fwd	GGAATTTAGGCTCCTTGCTG	57	30 sec	482
	Rev	TCCAAGGATCTATGTGCTACCC			
7	Fwd	TCCTGGAGGCTCTTGTTTTC	57	30 sec	506
	Rev	TTGGAATTGGTAAGGTTTGAAGA			
8	Fwd	CACCTGGCCTCGGTTATTAG	57	1 min	977
	Rev	TAAAAGCTGCCCAAGGTCAC			
9, 10, 11	Fwd	TGGGTGGCCAGATCATTATT	57	3 min	3108
	Rev	TTTGGAGAATCATGGCTTAGG			
12, 13	Fwd	AGCTTCAATCTTAAATAACTTGCAC	55	30 sec	599
	Rev	TTTCCCTACACAGCAGTAATAGC			
14	Fwd	TCCTATCACTATGTGAAGCCAGAA	61	30 sec	591
	Rev	AGCCTGGCCACAGAGTAAGA			
15, 16	Fwd	CAGCCTTCTTTTCTCATTCTCTTC	57	30 sec	722
	Rev	AGCCAGTTTTCCCACACATC			
17	Fwd	CCATTTGCCCTTCTGCTTTA	55	30 sec	459
	Rev	GCTTGATGGTATGGGCTTGT			
18	Fwd	TGGCCTTATCCTCATGCTTC	55	30 sec	425
	Rev	GTTGCAGTGAGCCAAGACTG			
19	Fwd	GCTTCTTTCCTAGTTCCTCTTCC	57	30 sec	491
	Rev	TCTTGGGTGAGTGAGATGAGAG			
20	Fwd	CATTGCATTTCGAGGGAATC	57	30 sec	433
	Rev	TGAATTCAGAAACTCCTGGACA			
21, 22	Fwd	TGCTTCAAGCCATCATTTTG	57	2 min	1506
	Rev	TTCATGAGCATCAAGGCAAA			
23	Fwd	GCATGCAGTTCTTGTTGCAT	58	30 sec	541

	Rev	CCAGCCTGAAATGCTGTTTT			
24	Fwd	TCCAAGAGGTATTCTGATGTGG	58	30 sec	592
	Rev	AAACATCGGAGGCACAATTC			
25	Fwd	TGGGACGAATGGGTGATTAT	58	30 sec	439
	Rev	TTCTGGGAACTAAATCTGTATGACC			
26	Fwd	TCATTCACTTTGTCCATTCCAG	58	30 sec	543
	Rev	TGTGCTTCAAAAGTGCTCCA			
27	Fwd	CTGTTGGTTGCATGCCTGT	58	30 sec	396
	Rev	GTGCCAAGTGTCAGCCATTA			
28	Fwd	CAAATGGACTCTCATCTTTTGC	57	30 sec	385
	Rev	TCATCCACATGGACTTTTGC			
29	Fwd	TGAATAAGTTGTGCCGTTGC	57	30 sec	574
	Rev	TGTGCAGAAGAATAAAACTGTTCA			
30	Fwd	GCACAGTAGTTTCCTCCTCTCA	57	30 sec	484
	Rev	CACACACTCCCCTTTGAGGT			
31	Fwd	TCTGGAAGATCCCCTTCATC	57	30 sec	529
	Rev	GACTGCTTAACCGCAAGCTC			
32	Fwd	GAGCTTGCGGTTAAGCAGTC	57	1 min	1062
	Rev	CCATAGAGCCCTTGAAGAATG			
33	Fwd	CATTGTGAGTACAGAGGAAACAGTC	57	1 min	692
	Rev	TCTAGCCTGCCACCTGTTTT			

* Exon 1 in NM_001165979.1

Table A.2: Primers for Sanger sequencing of *PLCE1* exons

Exon	Forward primer	Reverse primer
1	GGGAGCGGACTGTGAACG	GAGCGCGAGGACACTTTTC
2_1	TTTTGGCTGGAAGCAGAAGT	CACAGGTATGAGAACAGAAGCTG
2_2	GATCTACCACCTTAAAACCCTGA	AGGAAGGCCATGCTGATG
2_3	CACATACTGTCAGACGAAGTGG	-
2_4	CTGGAACTAGACAGACCTTCCA	-
2_5	TGCTTTGAAGGCTCTTGTGA	-
1*	CCTCCCTCGATTCTGGGTAT	AACAGTCCCCAGGATTCCAT
3	TTTGCACTTGGAGCATCTGA	-
4	CAGAACTCTTCACTAAGCAGAGGA	-
5	CCCAGCCAGGACCTACAG	-
6	GGAATTTAGGCTCCTTGCTG	TCCAAGGATCTATGTGCTACCC
7	TCCTGGAGGCTCTTGTTTTC	-
8_1	CACCTGGCCTCGGTTATTAG	-
8_2	GACGGAGCTCATCCCTTG	-
8_3	TGCTGGATTAAGTAGCCTGAC	-
9	TGGGTGGCCAGATCATTATT	-
10	CGGTCAGCCTTAATGTAGGTC	-
11	CCACCAGATTAGCCCATTCA	-
12, 13	AGCTTCAATCTTAAATAACTTGCAC	TTTCCCTACACAGCAGTAATAGC
14	TCCTATCACTATGTGAAGCCAGAA	-
15	CAGCCTTCTTTTCTCATTCTCTTC	ATGGCCCGTGAGGTAGGTAT
16	ATCCCTTGCAGAAGTTCGAG	AGCCAGTTTTCCCACACATC
17	CCATTTGCCCTTCTGCTTTA	-
18	TGGCCTTATCCTCATGCTTC	-
19	GCTTCTTTCCTAGTTCCTCTTCC	-
20	CATTGCATTTCGAGGGAATC	-
21	TGCTTCAAGCCATCATTTTG	-
22	TCTAGGAAAGCTGTTGGGACA	-
23	GCATGCAGTTCTTGTTGCAT	-
24	TCCAAGAGGTATTCTGATGTGG	AAACATCGGAGGCACAATTC
25	TGGGACGAATGGGTGATTAT	-
26	TCATTCACTTTGTCCATTCCAG	-
27	CTGTTGGTTGCATGCCTGT	-
28	CAAATGGACTCTCATCTTTTGC	-
29	TGAATAAGTTGTGCCGTTGC	-
30	GCACAGTAGTTTCCTCCTCTCA	-
31	TCTGGAAGATCCCCTTCATC	-
32_1	GAGCTTGCGGTTAAGCAGTC	-
32_2	AAAGTATCCAAACCAAGGAGGA	-

32_3	ATGCTATTGAACACCGCCTA	-
33_1	CATTGTGAGTACAGAGGAAACAGTC	-
33_2	TTGATGATTTCTGAACTGAAGC	-

* Exon 1 in NM_001165979.1

Multiple primers were used for some exons to ensure complete coverage

Table A.3: Primer and reporter sequences for Custom TaqMan SNP genotyping assays

Variant	Primer/ reporter	Sequence
COX-2 rs20417	Fwd	CCCCCTCCTTGTTTCTTGGAA
	Rev	TGCTTAGGACCAGTATTATGAGGAGAA
	Reporter 1	ACCTTTCCCGCCTCTC
	Reporter 2	ACCTTTCCCCCCTCTC
CASP8 rs1045485	Fwd	ACCACGACCTTTGAAGAGCTT
	Rev	TCCATGAGTTGGTAGATTTTCAAAATCTCA
	Reporter 1	CCCCACGATGACTG
	Reporter 2	CCCCACCATGACTG
ALDH2 rs441	Fwd	AGCCTGGGTGCCAGAGAGA
	Rev	CCTGACAGCATTCACTTAGAACAAC
	Reporter 1	CTCGGCCTCAAAA
	Reporter 2	ACTCGGTCTCAAAAA

Table A.4: Primers and PCR conditions for Sanger sequencing of somatic mutations

Sample	*Gene*	Primer	Primer sequence	Tm (°C)	Annealing temp (°C)	Extension time	Product size (bp)
386T- P1282	*MEF2C*	Fwd	CACATTTGAGACAAACAGCTCA	58.97	54	30 sec	297
		Rev	CCAATTTTATCAAAGCTACCACCT	59.81			
	MKL1	Fwd	CCATCTCACCAAAGGTGTCC	60.36	55	30 sec	378
		Rev	AGCTGAAGCAGGAGCTGAAG	60.04			
	PPM1D	Fwd	CCCCCTGATGAAGAAGCATA	60.03	56	30 sec	198
		Rev	CCAGATGCATTTCAGCAAAC	59.28			
	RBM26	Fwd	CTCCTGTTGTTGAAGGACCA	58.69	54	30 sec	304
		Rev	TTGGCACAGGCTACAAAGTG	59.9			
	TP53	Fwd	GAAGACCCAGGTCCAGATGA	60.05	56	30 sec	216
		Rev	ACTGACCGTGCAAGTCACAG	59.94			
T438-P1400	*APC* (112173713)	Fwd	GCTCAAGCTTGCCATCTCTT	59.72	54	30 sec	344
		Rev	TAGACCAATTCCGCGTTCTC	60.21			
	APC (112174228)	Fwd	AAGAAGCTCTGCTGCCCATA	60.12	54	30 sec	507
		Rev	TGTGGTTGGAACTTGAGGTG	59.57			
	APC (112174347)	Fwd	As above				
		Rev					
	ARHGAP21	Fwd	TGAAGAGAGGCTCAGGGAGT	59.13	54	30 sec	403
		Rev	TGAAAGGGGCATTTTAGCTG	60.2			
	ATAD5	Fwd	TTCATCCTCAAGAAGAGCATTT	58.07	54	30 sec	386
		Rev	CCCAGCCATAGCACAGATTT	60.1			
	CARS	Fwd	ATCCCGAGTCTCCTTCCAGT	60.07	54	30 sec	402
		Rev	GGTTGCAGCCATTTTCAGTAA	60.12			
	CNGB1	Fwd	CATCCTGCCCCTTGTTCTTA	60.07	62	30 sec	291
		Rev	TGAGCTAGGGGAAGTTGAGG	59.42			
	CORO2B	Fwd	GCATATGAGTGTGCCCTGAA	59.68	55		412
		Rev	GCAGCTGGAGTGAAGGTAGG	60.01			
	FCRL3	Fwd	TAGTCTCCAGGGTGGTGAGG	60.1	54	30 sec	417
		Rev	TGTCAGGGGAAAGACAACAA	59.11			
	KRT27	Fwd	AAGGCTTGAGACAACGTCAAA	59.91	62	30 sec	369
		Rev	AGCGAGACTCGGTCTCAAAA	60.13			
	MLL2	Fwd	ACATAGTTCTGGGCCCTCCT	59.96	64	30 sec	438
		Rev	GCGTGGTACTGATGCTTGTG	60.33			

	NLRC5	Fwd	GAGTGGGGTGTGGACAAGAT	59.82	54	30 sec	342
		Rev	AGAAAGGCCCAGAGAAGAGC	60.1			
	OR52A5	Fwd	GGTCACGAATTTGCTTGGTC	60.5	54	30 sec	461
		Rev	ACTCAGGGCTGCCATTCTTA	59.84			
	PMS1	Fwd	GGAACAAAACTTTGTCCTGACTG	60.06	54	30 sec	375
		Rev	TGTGTGGCCTGGTTCCTAAC	60.95			
	SP1	Fwd	GTGAATGCTGCTCAACTCTCC	60.01	58	30 sec	340
		Rev	CCAGAAGTGCTCCTTCCTCA	60.52			
	TP53	Fwd	GAAGACCCAGGTCCAGATGA	60.05	56	30 sec	216
		Rev	ACTGACCGTGCAAGTCACAG	59.94			
	ZNF521	Fwd	CCTCAAGTCAGGGTGCATTT	60.11	54	30 sec	368
		Rev	GGGCAGCCTACACTATGGAA	60.1			
	ZNF750	Fwd	CCAGCAGGTAAGGCGAGTAG	60.03	54	30 sec	308
		Rev	TTCTGCATTTGTTCCAGTCG	59.84			
T437-P1377	*RECQL4*	Fwd	CCAGCTCACTGAACTCCACA	60.02	55	30 sec	344
		Rev	GTTGGAGACGAGGTTGGAGA	60.24			
T441-P1116	*ARHGEF2*	Fwd	TAAGACCTGCAGGCATCTCC	60.36	57	30 sec	460
		Rev	AAGGCTCGGTATGACCACAG	60.13			
	COX6C	Fwd	GGGACAGTCACCTGTATTTGC	59.47	55	30 sec	343
		Rev	CCTCCACCCCTTAGGAAATG	60.67			
	FLT3	Fwd	AGCCTGCGTGGTAGAGTGA	57.89	57	30 sec	399
		Rev	TGCAGCATCTCCTGTAGCAA	50			
	FZD6	Fwd	AAGCAAAAGCTCGACCAGAA	45	57	30 sec	279
		Rev	CCAAACTTCAGTTGGCAAATC	42.86			
	IL21R	Fwd	CCCTCACCTTACCCTCATCC	60.7	55	30 sec	384
		Rev	GGTAGCCGTCATCCTCACAG	60.68			
	LARGE	Fwd	GGCAGCTGTATCAGAGCACA	55	57	30 sec	311
		Rev	GGTGTGCCTAGCTCTCCATC	60			
	PAX7	Fwd	ACCCCTGCCTAACCACATC	59.79	55	30 sec	321
		Rev	AACAATGGGAGGACATCTGG	59.78			
	TET2	Fwd	TTCCACAGGTTCCTCAGCTT	59.84	57	30 sec	358
		Rev	TGCTGGCAGTTGTCCTGTAG	60.05			
	TMPRSS2	Fwd	AAGATTCTGCCAACCTGCTT	58.94	55	30 sec	348
		Rev	CTCCCTGTGTGGTTTTTGGT	59.86			

T443-P1408	*DNAH10*	Fwd	AATCTCTTGACCGCAAAAGC	59.46	63	30 sec	408
		Rev	ATCTTCATCCCAACGACAGC	60.08			
	JUNB	Fwd	AGCCTGACAGGGCTTTTG	58.96	57	30 sec	431
		Rev	GAAGAGGCGAGCTTGAGAGA	59.97			
	KL	Fwd	CACTCAGGGAGGTCAGGTGT	60.15	63	30 sec	399
		Rev	CCTGAGACAAACCAGCCATT	60.11			
	OSBPL3	Fwd	TTGTCTCCCCCACAAGGTTA	60.34	63	30 sec	362
		Rev	AGGTCCCGAAACCTTTTTCT	59.08			
	RORA	Fwd	TCATTGTTTCCCCTCCTTTG	59.9	54	30 sec	356
		Rev	CCTGACGGTGTGTCCTTTCT	60.15			
	RTL1	Fwd	GTTTGGTCGTCGATTTGGAT	59.8	63	30 sec	422
		Rev	CGCCATCACAACGTCTACTG	60.32			
	SPRX	Fwd	CCCCAAGTTCGCTATTACCA	59.95	57	30 sec	368
		Rev	GGTTGAGAACAACCCAGGAA	59.94			
	VAMP4	Fwd	GCCCTGTTCTCACAGAGGTT	59.3	63	30 sec	407
		Rev	GCAAACTGATCTGCAAGCTG	59.75			
T442-P1406	*GPRASP2*	Fwd	GGTCCAAGTAATGGGTGGTG	60.09	54	30 sec	256
		Rev	TGATCCAGACTCAGCAGTGG	59.98			
	HPX	Fwd	GGTCTCACCAAATGCCTGTT	59.97	54	30 sec	307
		Rev	ACAAGCTCAGGGAAAGTGGA	59.84			
	WDR17	Fwd	TTCTAGACTCCTGCACAAAGTCA	59.18	54	30 sec	220
		Rev	CGAAGACATGCAAAACCAGA	59.84			
232T-P662	*MYH14*	Fwd	CCATCTCCCTCAAACTGGAA	60.04	54	30 sec	334
		Rev	TGATGGCTGGTTGTTTTTCA	60.09			
	TOPAZ1/ C3orf77	Fwd	GGGTCAGAGGGCAGTAGGTA	59.16	62	30 sec	321
		Rev	CTGATGCTGCAACCGACTTA	60.01			
T416-P1354	*CXorf30*	Fwd	ACAAACATCGACCCTGCATA	59	57	30 sec	237
		Rev	AACTTGGTTTGGTCCAAGAA	57.15			

Table A.5: Primers and PCR conditions for Sanger sequencing of somatic mutations in recurrently mutated genes

Sample	Gene	Chr location	Primer	Sequence	Tm (°C)	Annealing temp (°C)	Extension time	Product size (bp)
T443-P1408	*GPR98*	5:89953731	Fwd	TGCAAAAGTTTCATTTATCCAAAA	59.89	54	30 sec	307
			Rev	TTTCTCCCTGTCTGAACTCCA	59.83			
T441-P1116	*GPR98*	5:90046431	Fwd	AGGGAACCCCTTGTGACTTT	59.83	54	30 sec	263
			Rev	CCATAAGGCAGGATTTGGTC	59.39			
T438-P1400	*SRRM2*	16:2815904	Fwd	CCGAAAAATCGAGGTCTTCA	60.18	54	30 sec	350
			Rev	AGAACGCTTCCGACTGGTT	59.86			
T441-P1116	*SRRM2*	16:2809653	Fwd	GAGAATCCAGGGCAGGTACA	60.07	54	30 sec	306
			Rev	TCCATCCTGTTGCTTTAGCC	60.21			

Table A.6: *TP53* primers and conditions for PCR

Exon	Primer	Primer sequence	Annealing temp (°C)	Extension time	Product size (bp)
2+3	Fwd	CAGCCATTCTTTTCCTGCTC	55	30 sec	498
	Rev	GGGGACTGTAGATGGGTGAA			
4	Fwd	CTGGTAAGGACAAGGGTTGG	55	30 sec	495
	Rev	GCCAAAGGGTGAAGAGGAAT			
5+6	Fwd	GCCGTCTTCCAGTTGCTTTA	55	30 sec	557
	Rev	GGGAGGTCAAATAAGCAGCA			
7	Fwd	CGACAGAGCGAGATTCCATC	64	30 sec	284
	Rev	GGGTCAGAGGCAAGCAGA			
8+9	Fwd	CAAGGGTGGTTGGGAGTAGA	55	30 sec	545
	Rev	ACCAGGAGCCATTGTCTTTG			
10	Fwd	TGCATGTTGCTTTTGTACCG	55	30 sec	300
	Rev	GAAGGCAGGATGAGAATGGA			
11	Fwd	AAAGCATTGGTCAGGGAAAA	55	30 sec	349
	Rev	CCACAACAAAACACCAGTGC			

Table A.7: *PPM1D* primers and conditions for PCR

Exon	Primer	Forward primer	Annealing temp (°C)	Extension time	Product length (bp)
1	Fwd	GCTCGCTCCACTCGACTC	58	1 min	907
	Rev	ATACTTTGGTTGCGCTCTGG			
2	Fwd	TTGTTGCCATTTGTATCCTGA	62	30 sec	544
	Rev	CCTTGGGGCTCACTGAAATA			
3	Fwd	CCTCTCTGAACAGGAATTTTGG	54	30 sec	449
	Rev	GCATGGCTTCGATTAGGTTC			
4	Fwd	TCAAATGCTTTTCTGCGTCT	54	30 sec	431
	Rev	CAGATCAAGGCAAATGCAAA			
5+6	Fwd	AGCTTTGTTTGGGCCACAG	62	30 sec	674
	Rev	TTCTGGGCTACGAGATTCAAA			
7_1	Fwd	TGAATGCATACCCCGTTTTT	55	30 sec	581
	Rev	TCTTTCGCTGTGAGGTTGTG			
7_2	Fwd	CCAAATGAAAGCCCAAGAAA	55	30 sec	709
	Rev	CATCACTTTTCCAGTCTGCTT			
7_3	Fwd	TTGTGACAATAGGGCTAAATGTT	55	30 sec	659
	Rev	CAAATTCAGCAACATGAGGAA			
7_4	Fwd	TCCCAGACCAATGGCATTAT	54	30 sec	590
	Rev	GATCACGCCACTGCACTCTA			

Table A.8: Results of Immunochip case-control analysis showing SNPs with P<1x10^{-3}

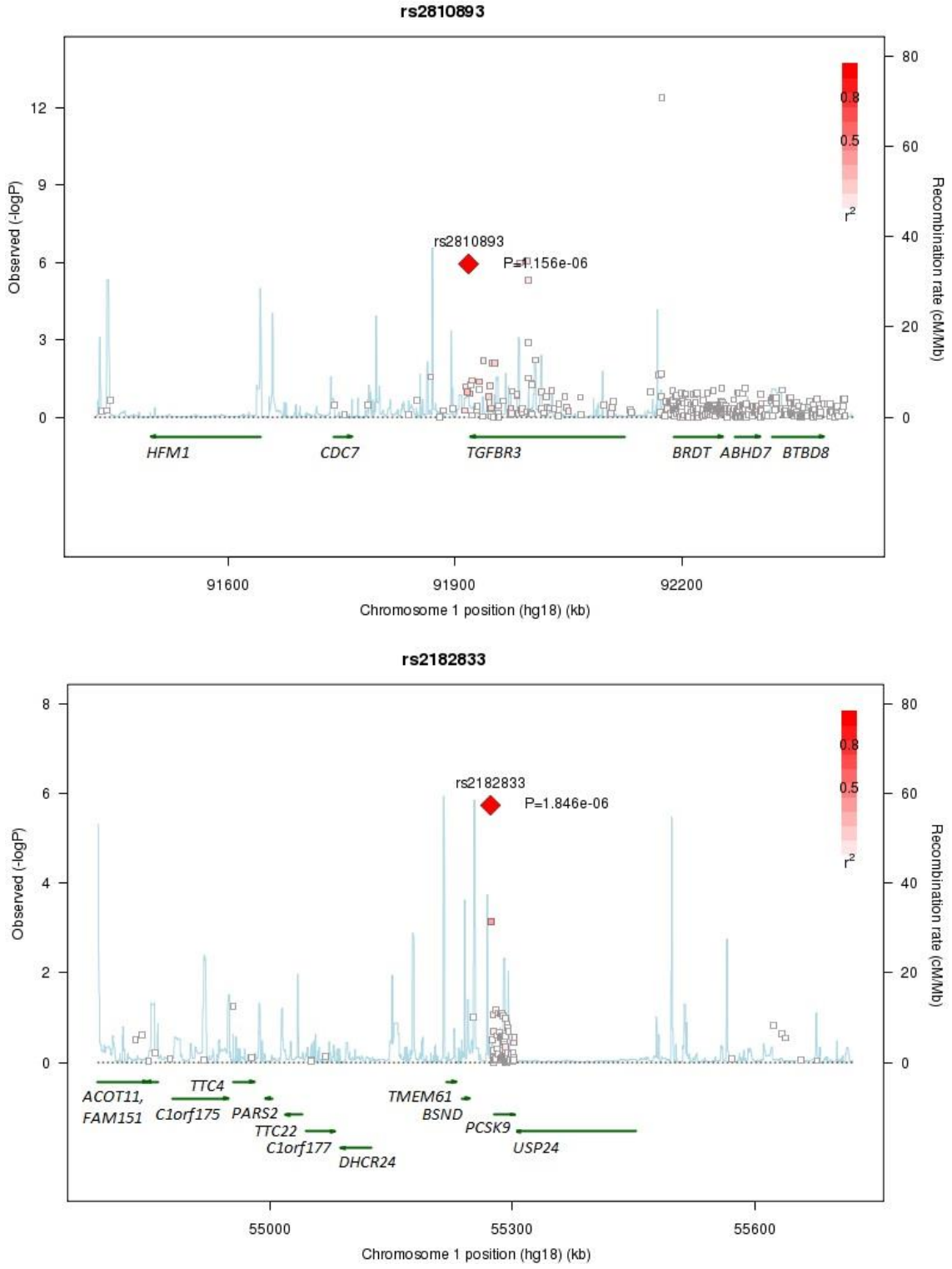
SNP ID	Chr	Position (b37)	Major / minor allele	MAF: Cases / controls	P-value	OR (95% CI)
rs9887787	1	92222143	G / A	0.060 / 0.152	8.86 x 10^{-7}	0.35 (0.23-0.54)
rs10493860	1	92212703	G / A	0.061 / 0.153	1.05 x 10^{-6}	0.36 (0.24-0.55)
rs2810893	1	92144970	G / A	0.239 / 0.377	1.16 x 10^{-6}	0.52 (0.40-0.68)
rs2182833	1	55500429	A / G	0.425 / 0.284	1.85 x 10^{-6}	1.86 (1.44-2.40)
rs11165441	1	92224347	G / A	0.059 / 0.143	4.96 x 10^{-6}	0.38 (0.25-0.58)
rs36590	22	30328070	G / A	0.022 / 0.081	9.51 x 10^{-6}	0.25 (0.13-0.48)
rs36596	22	30335269	G / A	0.022 / 0.081	9.51 x 10^{-6}	0.25 (0.13-0.48)
rs36600	22	30337586	G / A	0.022 / 0.081	9.51 x 10^{-6}	0.25 (0.13-0.48)
rs5763634	22	30350532	G / A	0.022 / 0.081	9.51 x 10^{-6}	0.25 (0.13-0.48)
rs4239932	22	30368384	C / A	0.022 / 0.081	9.51 x 10^{-6}	0.25 (0.13-0.48)
rs5763674	22	30386358	A / G	0.022 / 0.081	9.51 x 10^{-6}	0.25 (0.13-0.48)
rs5752993	22	30387160	A / G	0.022 / 0.081	1.01 x 10^{-5}	0.25 (0.13-0.49)
rs13147507	4	115334709	A / G	0.246 / 0.143	2.29 x 10^{-5}	1.96 (1.43-2.68)
rs13390918	2	199564895	G / A	0.480 / 0.353	2.59 x 10^{-5}	1.69 (1.32-2.17)
rs1400978	2	199565298	G / A	0.509 / 0.382	3.28 x 10^{-5}	1.68 (1.31-2.14)
rs12052337	2	181045431	A / G	0.182 / 0.288	3.89 x 10^{-5}	0.55 (0.41-0.73)
rs1001434	19	30205448	G / A	0.344 / 0.467	4.21 x 10^{-5}	0.60 (0.47-0.77)
rs228125	14	81338068	G / A	0.164 / 0.082	5.99 x 10^{-5}	2.18 (1.48-3.21)
rs1547354	21	26946709	G / A	0.016 / 0.063	6.64 x 10^{-5}	0.24 (0.11-0.51)
rs7714035	5	102644627	A / T	0.504 / 0.382	6.94 x 10^{-5}	1.64 (1.28-2.09)
rs12257237	10	6435157	C / G	0.305 / 0.202	0.000118	1.74 (1.31-2.30)
rs6043883	20	1683539	G / A	0.061 / 0.016	0.000138	4.09 (1.87-8.92)
rs6962404	7	7442169	A / C	0.230 / 0.139	0.000140	1.85 (1.34-2.54)
rs62226440	22	30590222	G / A	0.014 / 0.057	0.000146	0.24 (0.11-0.53)
rs17728461	22	30598552	G / C	0.014 / 0.057	0.000146	0.24 (0.11-0.53)
rs72660277	5	150181241	G / C	0.358 / 0.472	0.000150	0.62 (0.49-0.80)
rs13419935	2	199565637	G / C	0.079 / 0.153	0.000156	0.48 (0.32-0.70)
rs10176099	2	199566014	A / G	0.079 / 0.153	0.000156	0.48 (0.32-0.70)
rs9809480	3	60448671	A / G	0.161 / 0.084	0.000161	2.08 (1.41-3.06)
rs17643918	22	30273725	G / A	0.022 / 0.069	0.000176	0.30 (0.15-0.58)
rs9620926	22	30275091	G / A	0.022 / 0.069	0.000176	0.30 (0.15-0.58)
rs6739633	2	13028547	G / A	0.074 / 0.145	0.000176	0.47 (0.31-0.70)
rs2643468	5	96657330	G / A	0.425 / 0.539	0.000179	0.63 (0.49-0.80)
rs13386177	2	199566857	T / A	0.520 / 0.406	0.000196	1.58 (1.24-2.02)
rs923396	12	29122650	G / A	0.412 / 0.526	0.000204	0.63 (0.50-0.81)
rs41162	22	30408710	G / A	0.020 / 0.065	0.000218	0.29 (0.15-0.58)
rs3762315	1	67772011	A / G	0.206 / 0.303	0.000263	0.60 (0.45-0.79)
rs7816547	8	11147853	A / G	0.130 / 0.063	0.000264	2.21 (1.43-3.42)
rs3762316	1	67772023	A / G	0.207 / 0.304	0.000272	0.60 (0.45-0.79)
chr1:160854085	1	160854085	C / A	0.006 / 0.038	0.000289	0.14 (0.04-0.48)
rs3851229	6	111854660	A / G	0.065 / 0.020	0.000291	3.46 (1.70-7.05)

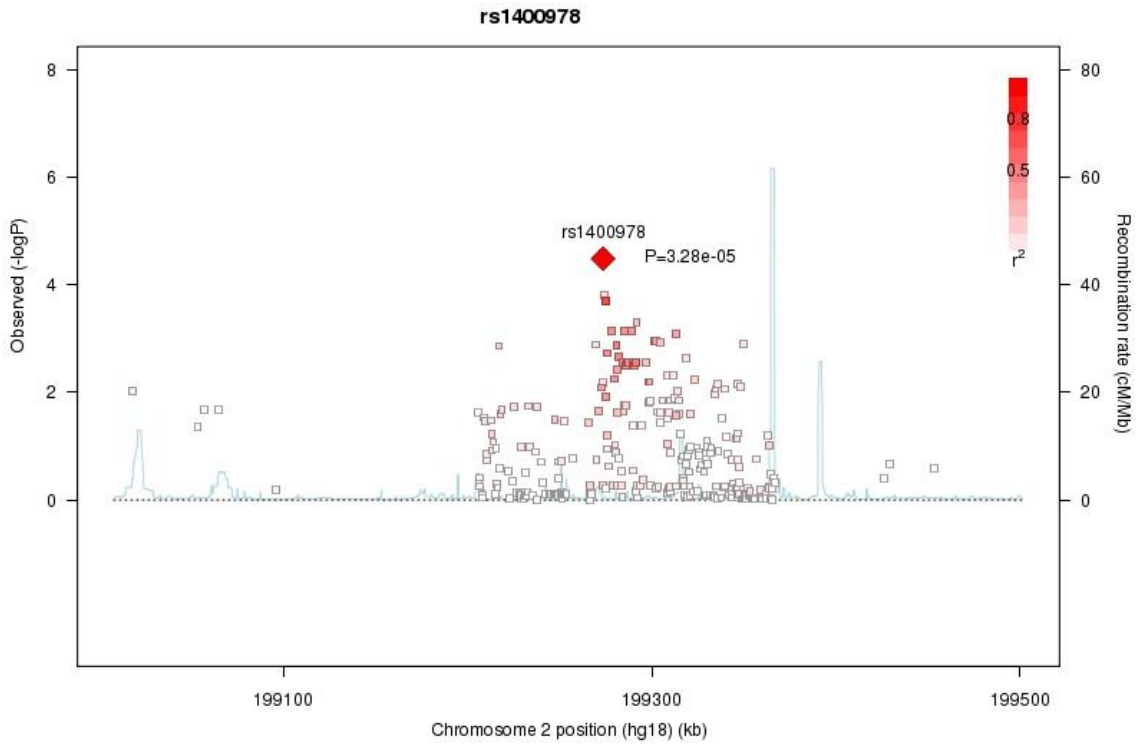
rs10139869	14	24346429	C / A	0.182 / 0.104	0.000311	1.91 (1.34-2.74)
rs62199977	2	186055665	C / G	0.034 / 0.086	0.000314	0.37 (0.22-0.65)
rs17103942	1	48410683	A / T	0.189 / 0.110	0.000318	1.89 (1.33-2.68)
rs8069115	17	40469180	G / A	0.149 / 0.235	0.000353	0.57 (0.42-0.78)
chr20:1554639	20	1554639	A / G	0.326 / 0.228	0.000359	1.64 (1.25-2.15)
rs2335813	3	21694884	A / G	0.212 / 0.308	0.000365	0.61 (0.46-0.80)
rs17106245	14	69206572	G / A	0.036 / 0.088	0.000368	0.39 (0.22-0.66)
rs17689022	16	68515208	G / A	0.135 / 0.069	0.000381	2.12 (1.39-3.22)
rs2834868	21	36632643	A / G	0.290 / 0.393	0.000387	0.63 (0.49-0.81)
rs7164531	15	88814356	A / G	0.389 / 0.496	0.000408	0.65 (0.51-0.82)
rs1800775	16	56995236	A / C	0.284 / 0.386	0.000410	0.63 (0.49-0.82)
chr8:57021505	8	57021505	A / G	0.029 / 0.077	0.000413	0.36 (0.20-0.65)
rs3130267	6	33306794	C / A	0.344 / 0.245	0.000440	1.61 (1.23-2.11)
rs6043636	20	1647384	G / A	0.085 / 0.033	0.000440	2.68 (1.52-4.73)
rs10818474	9	123489964	A / G	0.338 / 0.443	0.000442	0.64 (0.50-0.82)
rs10143801	14	88321793	A / G	0.372 / 0.478	0.000459	0.65 (0.51-0.83)
rs12338330	9	30753635	G / A	0.050 / 0.108	0.000467	0.44 (0.27-0.70)
rs2700986	7	37389804	G / A	0.110 / 0.051	0.000467	2.29 (1.43-3.69)
rs4134832	19	7691698	G / A	0.144 / 0.076	0.000480	2.03 (1.36-3.04)
rs8109557	19	55211859	G / A	0.216 / 0.310	0.000481	0.61 (0.47-0.81)
rs2760512	1	192561545	G / A	0.363 / 0.469	0.000487	0.65 (0.51-0.83)
rs13186205	5	40289988	G / A	0.040 / 0.092	0.000490	0.41 (0.24-0.68)
rs6864103	5	40302452	G / A	0.040 / 0.092	0.000490	0.41 (0.24-0.68)
rs17226632	5	40311005	A / G	0.040 / 0.092	0.000490	0.41 (0.24-0.68)
rs10462010	5	40311568	A / G	0.040 / 0.092	0.000490	0.41 (0.24-0.68)
rs4957127	5	40316010	T / A	0.040 / 0.092	0.000490	0.41 (0.24-0.68)
rs10858396	9	138275776	G / A	0.534 / 0.428	0.000496	1.54 (1.21-1.96)
rs11893706	2	199583546	A / G	0.155 / 0.239	0.000503	0.58 (0.43-0.79)
rs4747989	10	12590175	G / A	0.230 / 0.326	0.000507	0.62 (0.47-0.81)
rs8061192	16	9437050	A / C	0.254 / 0.167	0.000523	1.70 (1.26-2.30)
rs2491101	9	130215249	G / A	0.347 / 0.451	0.000536	0.65 (0.51-0.83)
rs12886388	14	81273539	A / G	0.192 / 0.116	0.000555	1.82 (1.29-2.57)
rs12259024	10	6411528	C / A	0.223 / 0.141	0.000569	1.75 (1.27-2.40)
rs5756708	22	37903578	G / A	0.407 / 0.512	0.000570	0.65 (0.51-0.83)
rs8128521	21	40452693	G / A	0.194 / 0.117	0.000572	1.82 (1.29-2.56)
rs1914011	3	2717559	A / C	0.112 / 0.186	0.000580	0.55 (0.39-0.77)
rs1885842	14	88335978	G / A	0.128 / 0.206	0.000594	0.56 (0.41-0.78)
rs12269345	10	102005620	G / A	0.209 / 0.300	0.000602	0.62 (0.47-0.81)
rs10105920	8	8648737	G / A	0.480 / 0.376	0.000604	1.53 (1.20-1.96)
rs6084626	20	4086072	G / A	0.173 / 0.259	0.000610	0.60 (0.44-0.80)
rs7636212	3	33025871	A / G	0.148 / 0.080	0.000615	1.98 (1.33-2.94)
rs7910961	10	6077796	G / A	0.175 / 0.102	0.000650	1.86 (1.30-2.67)
rs7305646	12	17264337	G / A	0.482 / 0.378	0.000650	1.53 (1.20-1.95)
rs61334194	2	33702486	G / C	0.270 / 0.182	0.000676	1.66 (1.24-2.22)

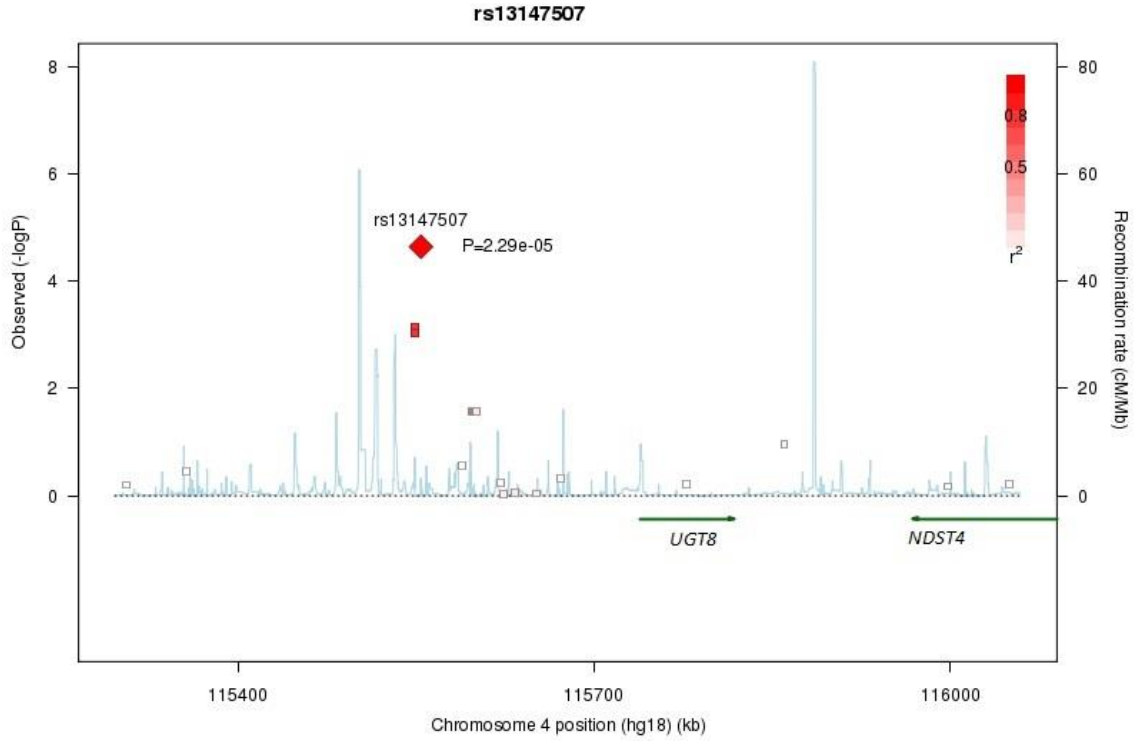
rs2332807	14	72710126	A / G	0.273 / 0.371	0.000676	0.64 (0.49-0.83)
rs9625919	22	30500958	C / A	0.018 / 0.057	0.000698	0.30 (0.15-0.63)
rs4337577	22	30512128	G / C	0.018 / 0.057	0.000698	0.30 (0.15-0.63)
rs4239933	22	30512414	G / C	0.018 / 0.057	0.000698	0.30 (0.15-0.63)
rs9625926	22	30539427	A / G	0.018 / 0.057	0.000698	0.30 (0.15-0.63)
rs9625933	22	30555216	G / A	0.018 / 0.057	0.000698	0.30 (0.15-0.63)
rs755856	8	10736552	G / C	0.142 / 0.077	0.000702	1.99 (1.33-2.98)
rs2542184	18	12749300	G / A	0.290 / 0.200	0.000703	1.63 (1.23-2.17)
rs12286448	11	79121342	A / C	0.281 / 0.192	0.000712	1.64 (1.23-2.19)
rs10788994	1	55500976	G / A	0.272 / 0.184	0.000717	1.65 (1.23-2.21)
rs11217040	11	118680648	C / A	0.272 / 0.184	0.000717	1.65 (1.23-2.21)
rs4446396	4	115329332	G / A	0.198 / 0.122	0.000720	1.78 (1.27-2.50)
rs10763790	10	30791355	G / C	0.095 / 0.165	0.000722	0.53 (0.37-0.77)
rs1356487	2	199569646	A / G	0.300 / 0.210	0.000726	1.62 (1.22-2.14)
rs1518099	2	199577125	A / G	0.300 / 0.210	0.000726	1.62 (1.22-2.14)
rs1878665	2	199580726	A / C	0.300 / 0.210	0.000726	1.62 (1.22-2.14)
chr15:79150671	15	79150671	A / G	0.177 / 0.263	0.000738	0.60 (0.44-0.81)
rs8010479	14	81125280	G / A	0.115 / 0.057	0.000762	2.16 (1.37-3.41)
rs2263657	20	1520233	C / A	0.115 / 0.057	0.000762	2.16 (1.37-3.41)
rs2263658	20	1520272	G / A	0.115 / 0.057	0.000762	2.16 (1.37-3.41)
rs2246154	20	1528038	C / A	0.115 / 0.057	0.000762	2.16 (1.37-3.41)
rs2250199	20	1539350	G / A	0.115 / 0.057	0.000762	2.16 (1.37-3.41)
rs11829361	12	9926680	A / G	0.004 / 0.029	0.000777	0.12 (0.03-0.52)
rs3116482	2	204553757	T / A	0.317 / 0.416	0.000778	0.65 (0.51-0.84)
rs2635357	4	118491393	G / A	0.210 / 0.300	0.000782	0.62 (0.47-0.82)
rs73207778	4	836762	G / A	0.040 / 0.008	0.000796	5.21 (1.78-15.2)
rs1400983	2	199605159	A / G	0.304 / 0.214	0.000803	1.61 (1.22-2.12)
rs3777723	6	167353701	G / A	0.394 / 0.296	0.000806	1.55 (1.20-1.99)
rs11680586	2	100746207	C / A	0.067 / 0.024	0.000807	2.96 (1.53-5.74)
rs6713524	2	100747357	A / G	0.067 / 0.024	0.000807	2.96 (1.53-5.74)
rs13000759	2	100753903	T / A	0.067 / 0.024	0.000807	2.96 (1.53-5.74)
rs7911500	10	6037726	G / A	0.076 / 0.029	0.000826	2.70 (1.48-4.93)
rs9295124	6	160570172	G / A	0.484 / 0.382	0.000845	1.51 (1.19-1.93)
rs6427405	1	157840353	G / A	0.103 / 0.173	0.000863	0.55 (0.38-0.78)
rs7713567	5	150430955	G / A	0.122 / 0.063	0.000864	2.08 (1.34-3.23)
rs2243603	20	1546911	C / G	0.122 / 0.063	0.000864	2.08 (1.34-3.23)
rs156029	5	131532634	A / G	0.259 / 0.353	0.000864	0.64 (0.49-0.83)
rs2583523	4	115928192	G / A	0.117 / 0.059	0.000885	2.12 (1.35-3.32)
rs718772	22	30504207	A / G	0.020 / 0.059	0.000887	0.32 (0.16-0.65)
rs11090598	22	30521770	A / G	0.020 / 0.059	0.000887	0.32 (0.16-0.65)
rs8135823	22	30522459	A / C	0.020 / 0.059	0.000887	0.32 (0.16-0.65)
rs8141765	22	30562239	C / A	0.020 / 0.059	0.000887	0.32 (0.16-0.65)
rs1978083	22	30570443	C / G	0.020 / 0.059	0.000887	0.32 (0.16-0.65)
rs10048885	22	30580252	A / G	0.020 / 0.059	0.000887	0.32 (0.16-0.65)

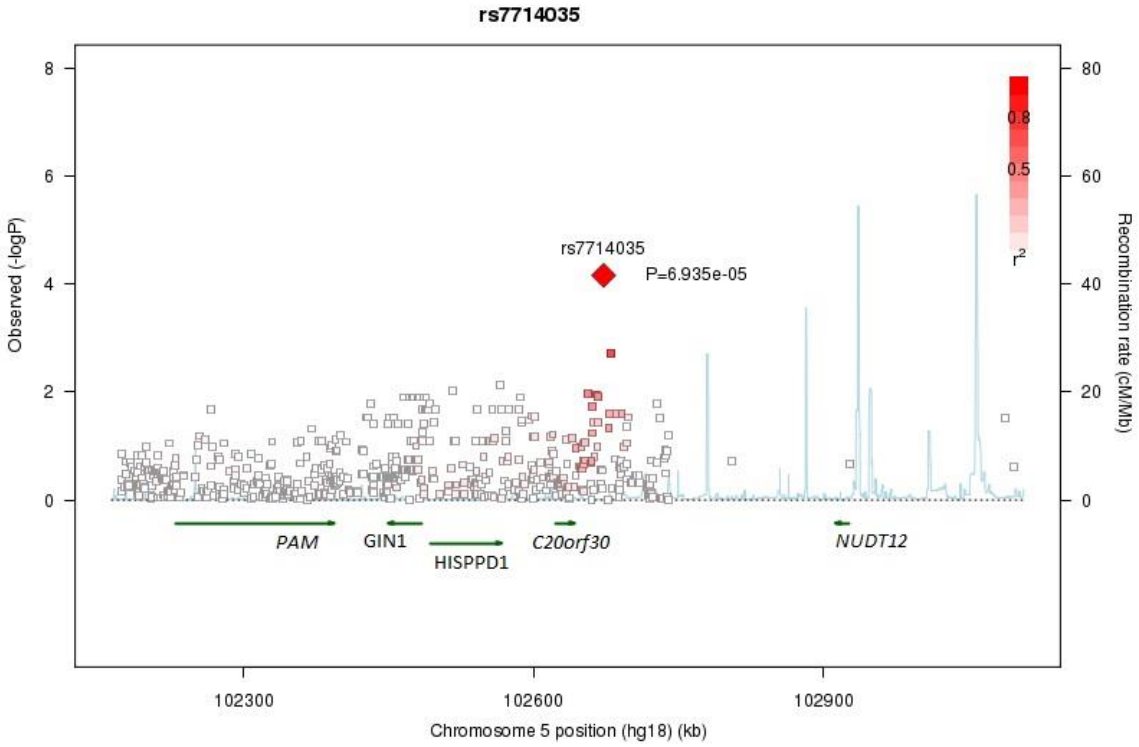
rs10771457	12	29127004	A / C	0.434 / 0.535	0.000887	0.66 (0.52-0.85)
rs62198554	2	186033255	T / A	0.016 / 0.053	0.000905	0.29 (0.14-0.63)
rs11121129	1	8268095	G / A	0.135 / 0.073	0.000913	1.99 (1.32-3.02)
rs1928042	13	47437216	A / C	0.473 / 0.373	0.000917	1.51 (1.18-1.93)
rs16944898	18	26124419	A / G	0.228 / 0.319	0.000919	0.63 (0.48-0.83)
rs40512	5	59805467	A / G	0.203 / 0.128	0.000919	1.75 (1.25-2.44)
rs4398410	3	189344011	G / A	0.353 / 0.259	0.000930	1.56 (1.2-2.030)
rs13138121	4	115329386	A / T	0.196 / 0.122	0.000934	1.76 (1.26-2.47)
rs4993442	22	30253256	A / C	0.011 / 0.043	0.000934	0.24 (0.10-0.60)
rs4823063	22	30308780	G / C	0.011 / 0.043	0.000934	0.24 (0.10-0.60)
rs36591	22	30328763	A / G	0.011 / 0.043	0.000934	0.24 (0.10-0.60)
rs5763644	22	30355676	A / G	0.011 / 0.043	0.000934	0.24 (0.10-0.60)
rs2117615	8	11153045	A / G	0.355 / 0.261	0.000935	1.56 (1.20-2.03)
rs9921632	16	75220263	A / G	0.369 / 0.469	0.000947	0.66 (0.52-0.85)
rs6985292	8	11140699	A / T	0.286 / 0.199	0.000952	1.61 (1.21-2.15)
rs7655600	4	81841042	G / A	0.074 / 0.135	0.000967	0.51 (0.34-0.76)
rs2237219	6	16375252	G / C	0.058 / 0.114	0.000984	0.48 (0.30-0.75)
rs2661548	4	48147774	G / A	0.124 / 0.198	0.000989	0.57 (0.41-0.80)

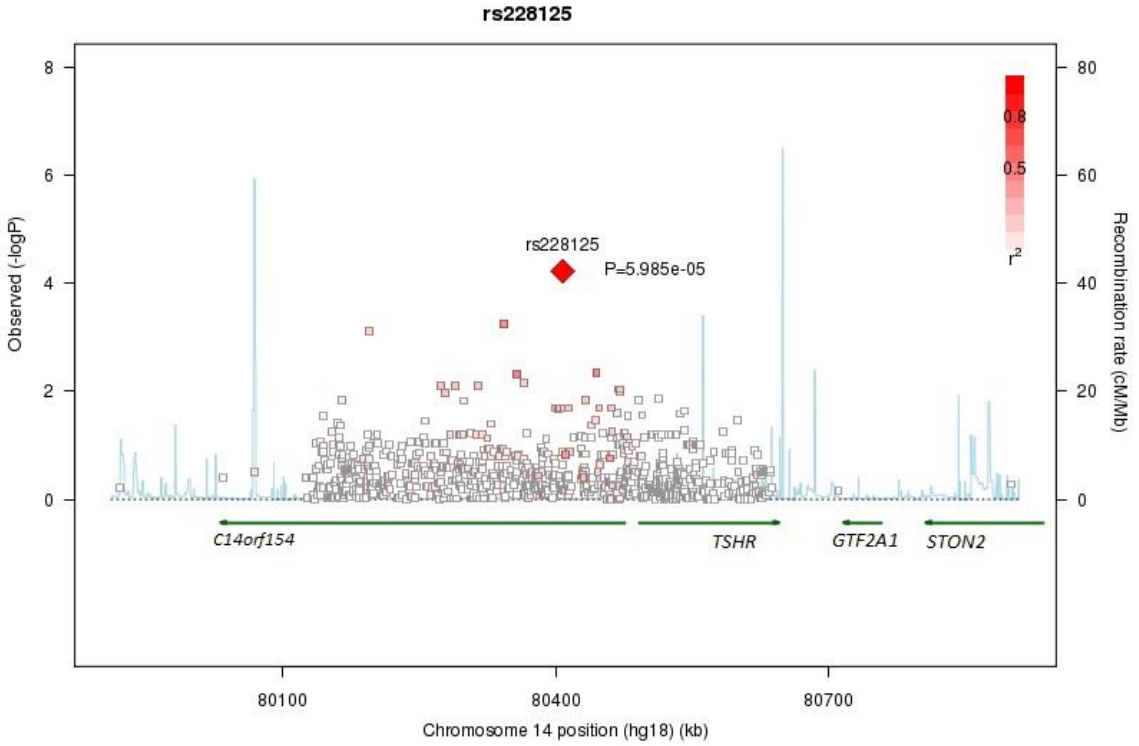
Figure A.1: SNAP association plots for variants with an Immunochip association of P< 1 x 10^{-4}

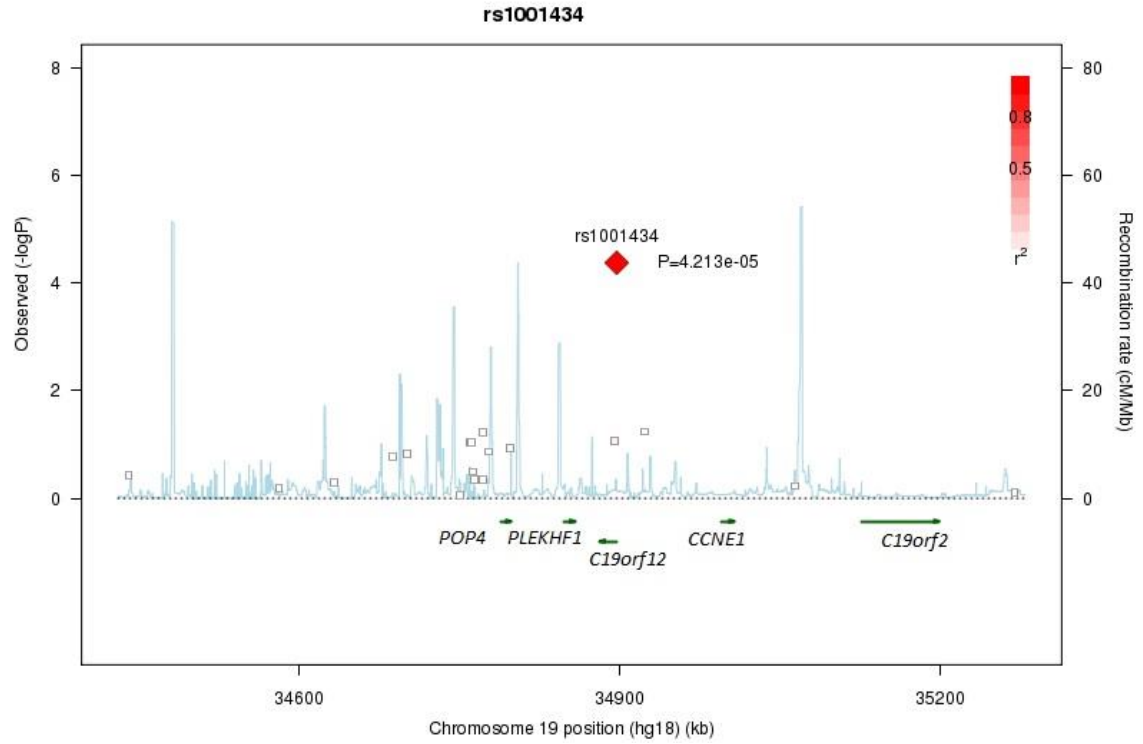


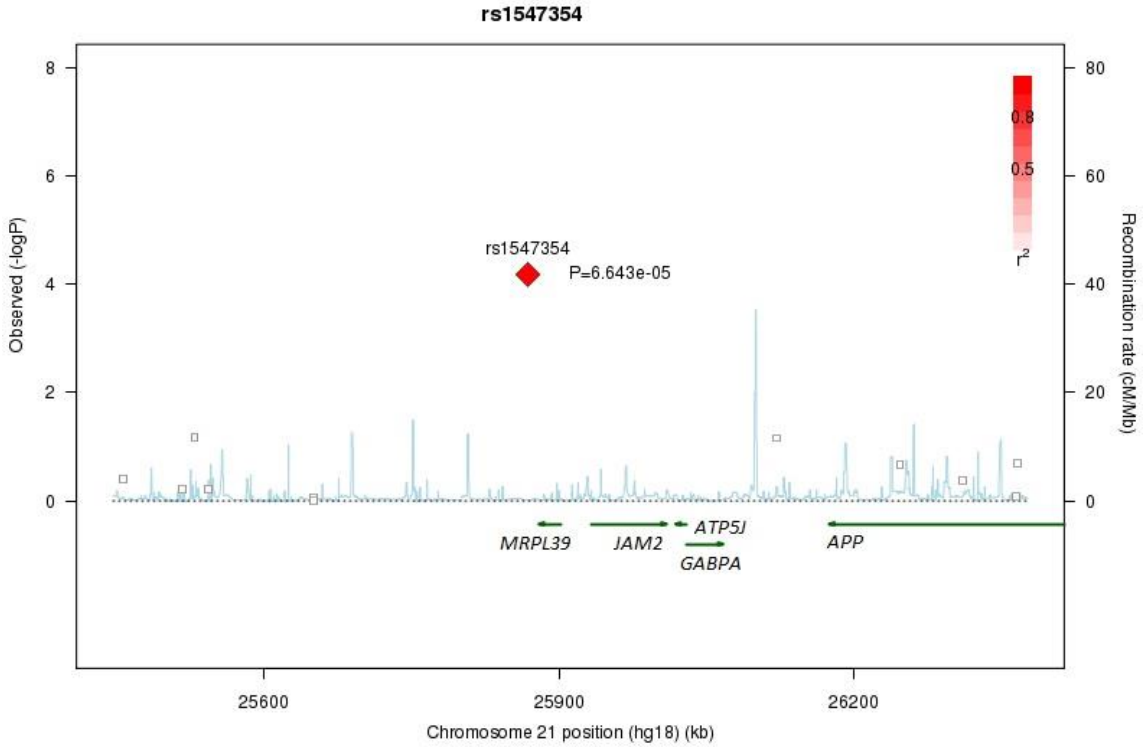

rs1400978
rs1400978
P=3.28e-05
0.8
0.5
r²
Observed (-logP)
Recombination rate (cM/Mb)
199100
199300
199500
Chromosome 2 position (hg18) (kb)


rs13147507
rs13147507
P=2.29e-05
0.8
0.5
r²
Observed (-logP)
Recombination rate (cM/Mb)
UGT8
NDST4
115400
115700
116000
Chromosome 4 position (hg18) (kb)


rs7714035
Observed (-logP)
Recombination rate (cM/Mb)
rs7714035
P=6.935e-05
0.8
0.5
r²
PAM
GIN1
HISPPD1
C20orf30
NUDT12
102300
102600
102900
Chromosome 5 position (hg18) (kb)


rs228125
Observed (-logP)
Recombination rate (cM/Mb)
rs228125
P=5.985e-05
0.8
0.5
r²
C14orf154
TSHR
GTF2A1
STON2
80100
80400
80700
Chromosome 14 position (hg18) (kb)


rs1001434
Observed (-logP)
Recombination rate (cM/Mb)
rs1001434
P=4.213e-05
0.8
0.5
r²
POP4
PLEKHF1
C19orf12
CCNE1
C19orf2
34600
34900
35200
Chromosome 19 position (hg18) (kb)


rs1547354
Observed (-logP)
Recombination rate (cM/Mb)
rs1547354
P=6.643e-05
0.8
0.5
r²
MRPL39
JAM2
ATP5J
GABPA
APP
25600
25900
26200
Chromosome 21 position (hg18) (kb)